Crossing Sentence Boundaries in Machine Translation

Thesis
presented to the Faculty of Arts and Social Sciencies
of the University of Zurich
for the degree of Doctor of Philosophy

by

Laura Mascarell

Accepted in the autumn semester 2017
on the recommendation of the doctoral committee:

Prof. Dr. Martin Volk (main supervisor)
Dr. Mark Fishel

Zurich, 2017

Abstract

Machine Translation systems translate sentences in a document independently of the discourse and any context information that crosses sentence boundaries. Often, the context provided in a sentence is not enough to correctly disambiguate a word, and the systems make incorrect lexical choices that negatively impact on the quality of the translations.

In this thesis, we attempt to integrate discourse knowledge into Machine Translation as a means to improve lexical choice in translation. Specifically, we develop discourse-aware methods for phrase-based Statistical Machine Translation systems, such as the sentence-level decoder Moses and the document-oriented decoder Docent. We also study the integration of discourse into Neural Machine Translation, whose high-quality translation output has recently attracted the attention of the Machine Translation community.

To improve the lexical choice of Machine Translation systems, our methods mostly focus on consistent translation of nouns and exploiting lexical chains, which are chains of semantically-related words in a document. Translation consistency, which consists of identifying a correct translation of a word and apply it consistently across the document, has been addressed in the literature with mixed results. In our experiments, we apply consistency in the translation of nouns in particular cases, where a consistent translation is expected, such as references to compounds and pairs of repeated nouns. In other experiments, we benefit from the semantic context provided by lexical chains in the source document to also keep the semantic similarity between words in the translation.

Zusammenfassung

Maschinelle Übersetzungssysteme übersetzen Sätze in einem Dokument unabhängig vom Diskurs und jeglichen Kontextinformationen, die über Satzgrenzen hinausgehen. Oft ist der durch einen Satz gegebene Kontext nicht ausreichend, um ein Wort korrekt zu disambiguieren, wodurch das System inkorrekte lexikalische Entscheidungen trifft, die die Qualität der Übersetzung negativ beeinflussen.

Diese Dissertation bezieht Diskursinformationen in die maschinelle Übersetzung mit ein, um lexikalische Entscheidungen während der Übersetzung zu verbessern. Hierzu werden Diskurs-sensible Methoden für Satz-basierte, statistisch-maschinelle Übersetzungssysteme, wie zum Beispiel den Satz-Level-Dekodierer Moses und den Dokumenten-Level-Dekodierer Docent, entwickelt. Auch wird die Integration von Diskursen in die neuronal-maschinelle Übersetzung untersucht, derenhochqualitative Übersetzungen jüngst die Aufmerksamkeit der Maschinellen Übersetzungs Community auf sich gezogen haben.

Um die lexikalischen Entscheidungen von maschinellen Übersetzungssystemen zu verbessern, fokussieren sich unsere Methoden auf die Übersetzungskonsistenz von Nomen und das Ausnutzen von lexikalischen Ketten. Dies sind Ketten semantisch verwandter Wörtern. Übersetzungskonsistenz, bestehend aus dem Identifizieren der korrekten Übersetzung eines Wortes und dem Anwenden dieser Übersetzung über das ganze Dokument hinweg, wurde in der Literatur mit gemischtem Erfolg adressiert. In unseren Experimenten wenden wir konsistente Übersetzungen von Nomen in bestimmten Fällen an, bei denen eine konsistente Übersetzung erwartet wird, wie etwa die Referenz von Komposita und Paare sich wiederholender Wörter. In anderen Experimenten erhalten wir die semantische Ähnlichkeit zwischen Wörtern in der Übersetzung durch den semantischen Kontext von lexikalischen Ketten im Quelldokument.

Acknowledgements

This thesis would have not been possible without the help and encouragement from several people, whom I would like to dedicate some words. Even though my Ph.D. studies at the University of Zurich started in November 2013, the first contact with Machine Translation occurred a while before.

My interest on Machine Translation started during the development of my Master thesis at Universtat Politècnica de Catalunya in Barcelona (Spain). There, I had the great privilege to work with *Meritxell González* and *Lluís Màrquez*, who supervised me on a topic related to the evaluation of Machine Translation systems. Their passion on the topic of Machine Translation became my source of inspiration and motivation, and thanks to their encouragement, I decided to continue my studies on the topic.

Soon, *Martin Volk* and *Mark Fishel* gave me the opportunity to do my Ph.D. studies in a wonderful country and supported me in all ways during the whole process. Whenever I struggled, they did not hesitate to get involved and give me the guidance and motivation I needed. I am especially grateful for the freedom they gave me to work on the experiments I was most interested in. Later, Mark continued his career path and settled in Tartu. Despite the distance, he remained as supportive as before, which I truly appreciate. The work on this thesis benefited also from *Andrei Popescu-Belis*' input and his great coordination of the MODERN project.

I would also like to thank all people who were part of the computational linguistics team during these years, since what I learned from each of them contributed to the development of this thesis. Especially, *Don Tuggener* and *Annette Rios*, who I consider my unofficial advisors, since they provided me with priceless advise and discussions that added great value to my experiments.

Sergi Oliva, who I met during my Master thesis, also provided great support throughout my whole Ph.D. studies. While he was realising his career dreams, he still managed to make me feel close to home every Sant Jordi, the most beautiful catalan tradition, which literally covers the city with books and roses.

I am grateful to my parents and to my brother for all their love. Also, my life in Switzerland would have not been the same without *Patrick Poullie*, who greatly contributed to my personal growth and happiness.

Last but not least, I would like to thank my grandmother, the most adorable and funniest woman I have ever met, and my grandfather, my teacher of life. He always encouraged me to be brave, to not fear setting my goals too high, and he showed me by example the

values of honesty, goodness, perseverance, and determination. And when I thought that there was nothing else I could learn from him, he showed me the most important lesson of all: true love lasts forever. This thesis is for both of you. Thanks for being part of my life.

Zurich, August 2017
Laura Mascarell

Contents

Abstract ii

Zusammenfassung iii

Acknowledgements iv

Contents vi

List of Figures xi

List of Tables xiii

1 Introduction 1
- 1.1 Research Questions . . . 2
- 1.2 Challenges in Document-level Machine Translation and its Evaluation . . . 3
- 1.3 Theoretical Background . . . 8
  - 1.3.1 Statistical Machine Translation . . . 8
    - 1.3.1.1 Phrase-based Statistical Machine Translation . . . 9
    - 1.3.1.2 Document-level Decoder Docent . . . 11
  - 1.3.2 Neural Machine Translation . . . 12
    - 1.3.2.1 Introduction to Artificial Neural Networks . . . 12
    - 1.3.2.2 Attention-based Neural Machine Translation . . . 14
  - 1.3.3 Neural versus Statistical Machine Translation . . . 15
  - 1.3.4 Technical Settings . . . 18
  - 1.3.5 Description of the Corpora used for the Experiments . . . 18
- 1.4 Relation to our own Published Work . . . 19
- 1.5 Thesis Outline . . . 21

2 Background on Discourse in Machine Translation 23
- 2.1 Cohesion in Translation . . . 23
- 2.2 Encouraging Lexical Consistency . . . 25
- 2.3 Discourse Connectives . . . 28
- 2.4 Pronominal Anaphora and Pronoun Translation . . . 29
- 2.5 Summary . . . 31

3 Analysis of Machine Translation Errors 33

3.1 Annotation of Discourse Translation Errors . . . 33
3.1.1 Annotation of Content Words: Nouns, Adjectives, and Verbs . . . 34
3.1.2 Annotation of Anaphoric Pronouns . . . 36
3.1.3 Annotation of Ambiguous Discourse Connectives . . . 38
3.2 Setup of the Spanish→English and German→French SMT systems . . . 39
3.3 Analysis of the Annotated Discourse Errors . . . 41
3.4 Summary . . . 49

4 Consistency, or No Consistency **51**
4.1 Introduction to Consistency . . . 52
4.2 Nominal References to Noun Compounds and Consistency . . . 52
4.2.1 Automatic Detection of References to Compounds . . . 53
4.2.2 Integration of the Consistency Method with the SMT System . . . 54
4.2.3 Consistent Translation of References to Compounds from German into French . . . 57
4.2.3.1 Evaluation of the Automatic Detection of References to German Compounds . . . 58
4.2.3.2 Manual Analysis of the Decode Approach . . . 60
4.2.4 Comparison of Approaches: Decode versus Post-editing . . . 64
4.3 Consistent Translation of Repeated Nouns . . . 66
4.3.1 Detection of Noun Pairs . . . 68
4.3.2 Classifiers for Consistent Translation . . . 69
4.3.2.1 Semantic Features . . . 70
4.3.2.2 Syntactic Features . . . 71
4.3.3 Analysis of the Classification Task . . . 74
4.4 Summary . . . 79

5 Exploiting Lexical Chains in SMT **81**
5.1 Introduction to Cohesion and Lexical Chains . . . 82
5.2 A Lexical Chain Model for SMT . . . 84
5.2.1 Building Source Lexical Chains . . . 84
5.2.2 The Lexical Chain Translation Model . . . 86
5.2.3 Computation of Semantic Similarity . . . 88
5.3 Setup of the Lexical Chain Translation Task . . . 90
5.4 Experimental Results . . . 91
5.5 Length, Density and Repetition in Lexical Chains . . . 96
5.6 Summary . . . 98

6 Document-level Neural Machine Translation **101**
6.1 Lexical Chains in Neural Machine Translation . . . 102
6.1.1 Computation of Semantic Similarity . . . 102
6.1.2 Annotation of Lexical Chains as Input Features . . . 103
6.2 Setup of the Word Sense Disambiguation Task . . . 104
6.2.1 Training and Test Corpora . . . 106
6.3 Evaluation of the Systems on the WSD task . . . 110
6.4 Summary . . . 111

7 Conclusions **113**

7.1 Future Research 117

Bibliography **119**

List of Figures

1.1 Sentence Translation in Phrase-Based Statistical Machine Translation . . 9
1.2 Image of a Brain Neuron. 13
1.3 Representation of an Artificial Neural Network. 14
1.4 Illustration of the Attention-based Neural Machine Translation Architecture. 15

3.1 Diagram of the Translation Error Annotation of Content Words. 34
3.2 Diagram of the Translation Error Annotation of Pronouns. 36
3.3 Percentage of the Relative Frequency of the Incorrect Translation of the German Pronoun *sie* into French. 44
3.4 Total of Fluency, Ambiguity, and Misalignment Error Annotations. 45
3.5 Total of Ambiguity Errors among Common Nouns, Adjectives, and Verbs. 46
3.6 Total of Local and Discourse Annotations for Ambiguous Common Nouns. 48

4.1 Example of Parse Tree Used to Obtain the Syntactic Features. 72
4.2 Translation Quality of the Maximum Entropy Classifier on the Test Sets. . 75

5.1 Example of Lexical Chains Detected with our Method. 86
5.2 Translation Examples of the Method. 94
5.3 Translation Examples of the Method on Consistency. 95

6.1 Pipeline to Obtain a Factored Corpus with Lexical Chains. 103
6.2 Examples of Contrastive Translations. 105
6.3 Accuracy of the Word Sense Prediction Grouped by Frequency of the Word Senses in the Training Set. 110

List of Tables

3.1 Annotated German Connectives. . . . 38
3.2 Annotated English Connectives. . . . 38
3.3 Data Sets Used for the Annotation Task 39
3.4 Total of Tokens of Each Annotation Test Set. . . . 40
3.5 BLEU Scores of the Annotation Test Sets 40
3.6 Overview of the Annotations per Error Category. . . . 42
3.7 Overview of Discourse and Local Annotations for English→Spanish. . . . 43
3.8 Overview of Discourse and Local Annotations for German→French. . . . 43

4.1 Consistency and Correctness of the `Cpd` System. . . . 62
4.2 Consistency and Correctness of the $\texttt{Cpd}_{split}$ System. . . . 62
4.3 Overall Percentages of Consistency and Correctness 64
4.4 Data to Train the German→French and Chinese→English Systems. . . . 65
4.5 BLEU Scores of the `Post-editing` and `Decode` Approaches. . . . 65
4.6 Comparison of the `Post-editing` and `Decode` Approaches with the Baseline. 66
4.7 Data to Build the German→English and Chinese→English Systems. . . . 67
4.8 Data to Train and Test the Classifiers. . . . 68
4.9 Example of Syntactic Features. . . . 71
4.10 Accuracy of the Classifiers on the Chinese Development Set. . . . 74
4.11 Accuracy of the Classifiers on the German Development Set. . . . 75
4.12 Accuracy of the Classifiers on the Chinese Test Set. . . . 76
4.13 Accuracy of the Classifiers on the German Test Set. . . . 76
4.14 Effects of Semantic Similarity on the Classification task. . . . 77

5.1 Data to Train the German→English System. . . . 91
5.2 Manual Evaluation of the Lexical Chain Results Compared to Using Lexical Resources. . . . 92
5.3 Weight Configurations of Length, Density, and Repetition. . . . 97
5.4 Manual Evaluation of the Study on Length, Density and Repetition in Lexical Chains. . . . 98

6.1 Word Sense Disambiguation Accuracy. . . . 107
6.2 Accuracy of the Word Sense Prediction by Frequency of the Word Senses in the Training Set. . . . 108
6.3 Translation Examples of the German→English System Trained on Lexical Chains. . . . 109
6.4 Average BLEU Scores on newstest2009-2013 111

Chapter 1

Introduction

Machine Translation (MT) is a field of computational linguistics that investigates the automatic translation of texts from one language to another. Its main goal is to produce high-quality translations comparable to human translations using MT systems. The task of translating a text is challenging for human translators, and even more for Machine Translation systems, as it requires an understanding of the entire document. While human translators consider the whole document as a unit, state-of-the-art MT systems translate each sentence individually, since isolated sentences are technically easier to handle. As a consequence, these systems ignore inter-sentential context information, and this discourse unawareness leads to incorrect lexical choice, negatively impacting on the translation quality of the MT system.

In example 1.1, extracted from a document of the alpine domain, the German noun *Träger* ("carriers" or "porters") appears in two different sentences. We observe that while in the human reference translation both occurrences appear translated into the same French noun *porteurs*, the Machine Translation system, which does not have discourse context, incorrectly translates them into *support* and *transporteur*.

(1.1) **Source:** Am 3. Juni schleppten Joe, Mac und ich die erste Traglast zum Lager II, während die ***Träger*** die unteren Lager mit Vorräten versorgten.
Am nächsten Morgen kamen die ***Träger*** unbegleitet vom Lager II zu uns herauf, als wir noch in den Schlafsäcken lagen.

Machine Translation: Le 3 Juin Joe, Mac, et j'ai traîné la première charge au camp II, tandis que le ***support*** fourni avec le roulement inférieur fournitures.
Le lendemain matin, le ***transporteur*** est arrivé seul à partir de Camp II à nous, car nous étions encore dans leurs sacs de couchage.

> **Human Reference:** Le 3, Joe, Mac et moi montâmes les premières charges au camp II, tandis que les ***porteurs*** faisaient la navette entre les camps inférieurs. Nous étions encore dans nos sacs de couchage, le lendemain matin, lorsque les ***porteurs*** arrivérent du camp II.

In this thesis, we investigate how to integrate discourse into MT systems in order to improve lexical choice and, consequently, the translation quality of the MT output. This research was supported from 2013 to 2016 by the Swiss National Science Foundation[1] under the Sinergia MODERN project (i.e. modelling discourse entities and relations for coherent machine translation),[2] a collaborative project between the following institutions: Utrecht Institute of Linguistics OTS, University of Geneva, Idiap Research Institute and University of Zurich. The general goals of this project are to assess discourse entities, such as noun phrases and pronouns, their relations, and to implement and integrate text-level features in Machine Translation. The MODERN project was a continuation of the Sinergia COMTIS project (2010-2013),[3] a Swiss collaborative project that focused on the translation of discourse connectives as a means to improve the coherence of MT output (Cartoni et al., 2011a), and it was divided into four sub-projects, one for each institution, to tackle different discourse-related problems.

In this chapter, we start by listing the research questions that guide the development of the experiments in this thesis (section 1.1). We then describe the challenges in the evaluation of lexical choice in MT systems (section 1.2), and next, we give an overview of the theoretical background necessary for the remaining chapters, such as a description of the most prominent MT approaches, the technical settings, and the data used for the experiments (section 1.3). Finally, we end this chapter by detailing the relation of the experiments in this thesis to our published work (section 1.4).

1.1 Research Questions

The goal of this thesis is to assess and integrate discourse knowledge into machine translation to improve lexical choice in the translation output. In the following, we list the research questions that we considered during the development of this thesis.

[1] http://www.snf.ch/en/Pages/default.aspx
[2] http://www.idiap.ch/project/modern
[3] http://www.idiap.ch/project/comtis

Research question 1: *How important is the use of discourse knowledge to improve lexical choice compared to the local context provided by the surrounding words?* This research question focuses on analysing to what extend discourse context or local context is needed to improve the lexical choice of Machine Translation systems. Specifically, we want to evaluate whether discourse-knowledge will improve the quality of the systems.

Research question 2: *What kind of inter-sentential context information is useful to improve lexical choice in Machine Translation, and how can it be integrated?* In particular, we aim at assessing what aspects of discourse entities and discourse relations from different text genres can improve the translation output, what kind of discourse-aware solutions can be integrated in different Machine Translation approaches, such as sentence-level and document-level decoders, and how they perform.

Research question 3: *Is translation consistency desirable in the output of Statistical Machine Translation?* Some researchers address this question in the literature. However, it is not clear whether consistent translations are the result of a better lexical choice, or whether more lexical variability should be introduced in translation.

Research question 4: *Can Neural Machine Translation benefit from discourse context, and how can it be integrated?* This question arises because Neural Machine Translation emerged as a new Machine Translation paradigm in 2016, but it only considers sentence-level context. At the time we addressed this research question, there was no published study on whether NMT systems could benefit from discourse, or how to integrate inter-sentential context.

1.2 Challenges in Document-level Machine Translation and its Evaluation

Machine Translation aims at generating grammatically and semantically correct translations, and in order to evaluate the quality of the automatic translations, we often compare them to the translations produced by humans, also called *human references*. Intuitively, the more similar the MT output is to the human reference, the better its translation quality.

In example 1.2, we observe that the German noun *Bericht* ("report", "story") occurs three times in the source text, one of them as a part of the compound *Jahresbericht*. The human reference uses the French translation *rapport* for all of them as they refer to the sense of *report*. However, the Machine Translation system translates *der Bericht* into *le récit* ("the story"), which is in the wrong sense. In this particular example, we show

that when the translation of *Bericht* does not match the one given by the reference, it indicates a translation error.

(1.2) **Source:** Mehr Details zum ***Jahresbericht*** Ausführlichere Informationen zur Jahresrechnung, ***Berichte*** über Aktivitäten und Projekte und diverse Statistiken des SAC sind in einem eigenständigen Jahresbericht publiziert ...
der ***Bericht*** steht auf www.sac-cas.ch unter der rubrik Downloads zur verfügung.

Machine Translation: Plus de détails au ***rapport annuel*** des informations exhaustives sur les comptes annuels, les ***rapports*** à des activités et projets et divers statistiques du CAS ...
le ***récit*** se trouve à télécharger sur www.sac-cas.ch sous la rubrique à disposition.

Human Reference: Pour davantage de détails sur le ***rapport annuel*** un ***rapport*** complet publié séparément donne des informations exhaustives sur les comptes annuels ...
ce ***rapport*** peut être téléchargé sur le site www.sac-cas.ch, rubrique «téléchargements».

To evaluate the performance of a translation system, we often use automatic metrics, such as BLEU (Papineni et al., 2002), which is widely used in the Machine Translation community to provide a reference score of a system's performance. This metric measures how similar a Machine Translation's output is to a reference translation (or references) by considering their n-gram overlap. BLEU has shown to correlate well with human evaluation, but this n-gram overlap approach has some limitations. Basically, if a correct translation of a term does not strictly match the one proposed by the references, the system is penalised. The more references we have, the better approximation on the translation quality we can get, as different translations can have equivalent meanings. Unfortunately, reference translations are expensive to produce, and we usually deal with only one.

In the following, we give some examples of the difficulties presented in Machine Translation when we need to compare their translations to a single human reference. All the examples are extracted from a German-French parallel corpus of essays on the alpine domain (see a more detailed description of the corpus in section 1.3.5).

When only one reference translation is available, we run the risk that correctly translated words do not match. In example 1.3, the MT system uses two different French translations, *chemine* and *sentier*, for the German noun *Weg* ("path"). None of them

match the translations given by the human reference *accès* and *itinéraire*, which are also different between them even though both *Weg* refer to the same path (i.e. *Brittanniahütte*). In contrast to examples 1.2, here, both reference translations (and also the MT translations) are in the same sense. Therefore, for stylistic reasons, the human translator may decide to introduce variability in the translation. However, this lexical variability is very challenging to reproduce for Machine Translation systems, and unless they generate exactly the same translations for *Weg* as in the reference, they are penalised by BLEU scores.

(1.3) **Source:** Neuer ***Weg*** Britanniahütte
Der neue ***Weg*** ist ausgeschildert, die Strecke ist ein wenig länger.

Machine Translation: Nouveau ***chemin*** de britannia
Le nouveau ***sentier*** est balisé, la route est un peu plus long.

Human Reference: Nouvel ***accès*** pour la Britanniahütte
Le nouvel ***itinéraire***, un peu plus long (1h 15), est balisé.

Similarly, in example 1.4 the German *Spitze* ("peak" of a mountain in this context) is translated into *cime* and *sommet* by the reference, and *pointe* by the MT system. The translation into *pointe* is not totally incorrect, although *cime* and *sommet* are better translations, since they clearly refer to the peak of a mountain. However, if the system uses only one of the reference translations for all ocurrences, it is also penalised, as one of them still does not strictly match the reference.

(1.4) **Source:** Bereits vier Tage nachdem Maurice Koechlin das Konzessionsgesuch eingereicht hatte, orientierte «Der Bund» seine Leserschaft in einem ausführlichen Frontartikel: Wenn man an den unverhofften Erfolg der Rigibahn und der Pilatusbahn denkt, sowie an die bedeutende Zahl von Touristen, welche jeden Sommer auf die ***Spitze*** dieser Berge fahren, so kann man keinen Augenblick zweifeln an einem noch grösseren Erfolg der Eisenbahn auf die ***Spitze*** der Jungfrau (4167m).
Machine Translation: Quatre jours après déjà Maurice Koechlin le Konzessionsgesuch transmises, nullement «la confédération» ses lecteurs dans un rapport Frontartikel: si on les unverhofften succès du funiculaire du Rigi et du Pilatusbahn pense, ainsi qu'à nombre de touristes qui, chaque été, à la ***pointe*** de ces montagnes, on ne peut pas douter un instant du succés plus grand encore sur le chemin de fer, à la ***pointe*** de la Jungfrau (4167 m).
Human Reference: Quatre jours seulement après l'octroi de la concession à

Maurice Koechlin, un éditorial du bund fait à ses lecteurs l'apologie du projet: si l'on pense au succès inespéré des chemins de fer du Rigi et du Pilate, ainsi qu'au nombre de touristes qui se rendent chaque été au ***sommet*** de ces montagnes, on ne peut douter un instant du succès plus grand encore du train menant à la ***cime*** de la Jungfrau (4167 m).

In example 1.5, the reference again uses two different French translations *époque* and *temps* for the German *Zeit* ("time"). Interestingly, the MT system uses the same French translations, but in different order. A human should judge whether both translations can be interchangeable in the context of each occurrence of *Zeit* in the document. In those cases where they can, the system is correct, but penalised if it does not match the reference.

(1.5) **Source:** Mit 8 Bildern (117-124) Vorbei sind die ***Zeiten*** des besinnlichen Dahingleitens über sanft gewellte, lang ausgezogene Hänge, verblichen der Ruhm des Blindenhorns, der Fuorcla Ziteil, des Hahnenmooses und des Hessisbohls, die es dem noch wenig schwungerpichten Skiläufer alten Stils gestatteten, kilometerweit hindernislos gemächlich abzufahren ...
Dieser Skistil ist Ausdruck seiner ***Zeit***.

Machine Translation: Avec 8 illustrations (117-124) l'***époque*** du Dahingleitens saisi de gewellte, longues pentes étaient pâli, la gloire du Blindenhorns, la Fuorcla Ziteil, du Hahnenmooses et du Hessisbohls, encore peu schwungerpichten qu'elle présente, vieux de plusieurs skieurs nous hindernislos tranquillement ...
Cette Skistil est l'expression de son ***temps***.

Human Reference: Avec 8 illustrations (117-121) les ***temps*** sont révolus des glissades tranquilles sur les longues pentes doucement ondulées. Le Blindenhorn, la Fuorcla Ziteil, le Hahnenmoos et le Hessisbohl, où le skieur de la vieille école, encore peu féru de vitesse, pouvait descendre paisiblement sur des kilomètres et des kilomètres sans rencontrer d'obstacles, ont vu pâlir leur renommée ...
Ce style du ski est l'expression de son ***époque***.

In other cases, as in example 1.6 and 1.7, the reference consistently translates the same term, and the Machine Translation system produces also a correct translation (consistent in example 1.6 and inconsistent in 1.7), but different from the reference. As a result, these correct translations are not positively reflected in the evaluation scores.

(1.6) **Source:** Fern vom Benzingeruch verfliegt mein «Rausch», und ich strebe mit ausdauerndem Schritt und mit klarem Kopf der ***Hütte*** zu ...
In dieser ***Hütte*** halten sich auch zwei französische Alpinisten auf, welche die «Haute-Route» der Dolomiten begehen.

Machine Translation: Loin du secteur trouver mon «ivresse», et je dernier avec ausdauerndem pas et clair, avec la tête vers la ***cabane*** ...
Dans cette ***cabane*** se tiennent aussi, deux grimpeurs français, à la «haute route» du tyrol du sud.

Human Reference: Loin de l'odeur de benzine, ma griserie se dissipe, et je peux approcher du ***refuge*** d'un pas ferme et l'oeil clair ...
Dans ce ***refuge***, deux alpinistes français qui suivent la haute route des dolomites.

(1.7) **Source:** Wir haben uns aus verschiedenen Gründen entschieden, die ***Touren*** zu veröffentlichen: Beide ***Touren*** wurden vom Autor und der Fachstelle Naturschutz/Natursport des SAC überprüft.
Machine Translation: Nous avons opté pour différentes raisons, les ***excursions*** en VTT: les deux ***courses*** ont été de l'auteur et de l'environnement Naturschutz/Natursport du CAS.
Human Reference: Nous avons décidé de publier ces suggestions de ***courses*** pour diverses raisons: les ***courses*** ont été examinées par l'auteur et par les responsables du secteur sport et environnement du CAS.

Finally, in some cases, the translation of the word does not appear in the translation, which might be due to several reasons, such as the translator decides to paraphrase the sentence for stylistic reasons; the translator substitutes the translation of a word with a pronoun to avoid repetition; or the reference is not a direct translation of the source text and, even though, source and reference have the same meaning, they are not strictly the same sentences. In such cases, we do not have a reference translation of a specific word to compare with the one generated by our system, as for the second occurrence of *Abgrund* ("abyss") in the example 1.8.

(1.8) **Source:** Der Mann über dem ***Abgrund*** rettet sich selbst der Mann über dem ***Abgrund*** gibt sich nicht auf. Er verzweifelt nicht.
Machine Translation: L'homme au-dessus du ***gouffre*** se sauve tout de même l'homme au-dessus du ***vide***, s'il n'y a pas à pas désespéré.
Human Reference: L'homme suspendu au-dessus du ***gouffre*** se sauve tout seul l'homme cependant ne perd pas courage.

The examples listed in this section give an overview of some of the problems we find when evaluating the translation quality of our systems compared to a reference translation. In the experiments described in this thesis, we use the automatic metric BLEU to compute the performance of our methods, but we also carry out manual evaluations of the output to get a better insight into the translation quality of our systems.

1.3 Theoretical Background

In this section, we give an overview of some technical aspects necessary to follow the remaining of this thesis. We first focus on the most prominent Machine Translation approaches: phrase-based Statistical Machine Translation (section 1.3.1), including sentence- and document-oriented decoders, and the new Neural Machine Translation approach (section 1.3.2). We dedicate most of the thesis to phrase-based SMT, as it was the state-of-the-art approach for high-resource language pairs. Recently, NMT outperformed phrase-based SMT for a number of languages pairs, and so, we attempt to integrate discourse knowledge in NMT systems in the last chapter of this thesis.

After the technical background on Statistical and Neural MT, we continue with a comparison of their strengths and weaknesses according to the literature (section 1.3.3). At the end of this section, we describe the technical settings of the systems we built (section 1.3.4) and the data used for the experiments (section 1.3.5).

1.3.1 Statistical Machine Translation

Statistical Machine Translation (SMT) is a Machine Translation approach, which produces translations based on statistical models. There are several approaches to SMT, such as phrase-based (Koehn et al., 2003), hierarchical (Chiang, 2005), and n-gram-based SMT (Mariño et al., 2006), which differ mostly in the way they handle the input data.

We briefly describe here the traditional phrase-based SMT approach, which translates sentences independently of each other, and a document-level approach, which was recently developed to handle document-level features. For a more detailed description of the algorithms, we refer to the *Statistical Machine Translation* book by Koehn et al. (2003) and Hardmeier et al. (2012)'s conference paper, respectively.

1.3.1.1 Phrase-based Statistical Machine Translation

Machine Translation systems based on phrase-based models show the best performance among all SMT approaches. Indeed, phrase-based SMT systems achieved the best translation results in a number of language pairs before the recent introduction of Neural MT.

Given an input sentence to translate, phrase-based models segment the sentences into short word sequences called phrases (or n-grams), which are independently translated into target phrases and reordered if the target language follows different word-order rules than the source language. The term used to define these word sequences (i.e. phrases) does not correspond to the linguistic concept of phrase, which represents a single meaningful unit, but to any non-overlapping sequence of words. This way, for example, the model can better translate the German preposition *von* ("of", "from", "by") when it is part of the sequence *reden von* into the English *talk about*. Figure 1.1 shows an example of a German sentence segmented into phrases and translated into English. In the English translation, the subject and the verb need to be reordered, since the word-order rules are different.

The translation of each phrase is, of course, not generated out of the blue. We use large parallel corpora (i.e. corpora in two different languages) to allow the system to learn how to translate phrases from one language to the other. In the training process, the system first aligns words from the source to the their counterpart on the target side for each parallel sentences and then extracts the phrase pairs. The source phrases, which are of different lengths (usually up to five words), are stored with their counterpart target phrases in a phrase table. This table maps the source phrases to their translations and the corresponding translation probabilities. It is important to consider both shorter and longer phrases. While longer phrases capture much more local context than the shorter ones, the latter occur more frequently, allowing the system to translate words that do not occur in the longer phrases.

In addition, the more parallel data we can use for training, the better translations we can obtain, as the system cannot learn translations that do not occur in the training data. For this reason, phrase-based SMT systems perform poorly with low-resource language

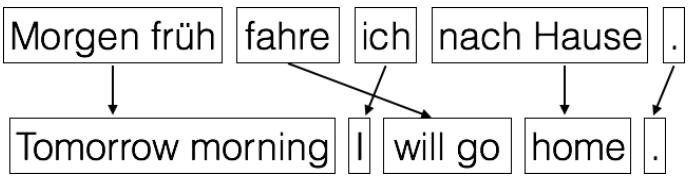


FIGURE 1.1: Translation of a sentence in phrase-based SMT

pairs, which need to rely on purely rule-based approaches or a hybridisation of rule-based and statistical models, as Rios Gonzales and Göhring (2013) propose for the translation from Spanish into Cuzco Quechua.

Once the system learned all possible translations of each phrase in the corpora, it is able to generate a vast amount of translation hypotheses for a given input sentence, but it needs to find the best translation among all them. To do so, phrase-based systems contain a set of feature functions (or models) that model different aspects of translation quality. Each of these models, which is trained independently of the others, gives a partial score to a translation hypothesis, and the goal of the system is to find the translation that maximises the linear combination of all partial scores. Phrase-based SMT systems can include a different variety of models to evaluate different aspects of the translation. In the following, we list the basic three models that are used to train virtually any phrase-based SMT system:

- The phrase translation model: This model uses the phrase translation table to obtain the probability of an input phrase being translated into a specific target phrase.
- The reordering model: The reordering model computes a probability of the order of the phrases in the translation hypothesis.
- The language model: This is a word-gram model (usually trigram to five-gram) built upon data from the target language. The model gives a probability to a target language word given the history of $n-1$ target words. This way, in a bigram language model, *does* is more likely to follow *he* or *she* than *they*.

More formally, given a source sentence s to translate, the overall score $f(s,t)$ of its hypothesis t is the linear combination of the partial scores produced by each model (e.g. phrase translation, reordering, and language model) h_k as in the equation:[4]

$$f(s,t) = \sum_k \lambda_k h_k(s,t), \tag{1.1}$$

where each feature function h_k has a weight λ_k, whose value represents the importance of the feature in the overall score, obtained with an optimisation technique such as MERT (Och, 2003a).

[4] In practice, the system internally uses logs to compute the equation 1.1.

The task of finding the best translation hypothesis, among all translations is, however, a NP-complete problem, since there is an exponential number of choices and it is computationally very expensive to explore all possible translations (Knight, 1999). To reduce this search space, phrase-based SMT systems use hypothesis recombination (Och et al., 2001) and the stack decoding algorithm (Koehn et al., 2003).

Hypothesis recombination is a dynamic programming technique that discards the translation hypotheses that can be reached by another path with a higher probability score. In other words, different phrase segmentations of the input sentence can lead to the same translation at different costs. In that case, the worse-scoring hypothesis cannot be the best translation, and it is therefore discarded. This technique makes the search more efficient, but does not solve the complexity issue. Therefore, phrase-based SMT systems implement the stack decoding algorithm, which reduces the search space by pruning out the bad hypotheses early on. Specifically, the algorithm distributes the phrase-translation hypotheses among stacks based on the number of words translated. That is, we find all phrase-translation hypotheses that translate one word in one stack, all the ones that translate two words in another stack, and so on. Then, if a stack gets too large, the algorithm discards the worse hypotheses in that particular stack.

In chapter 4 we perform several experiments to tackle translation consistency using phrase-based SMT systems. Since these systems translate one sentence at a time, we cannot model dependencies that cross sentence boundaries. Therefore, we store and pass the information that we need from the already translated sentences to the next ones. Another way to handle long-range dependencies would consist on removing the sentence boundaries and treat the whole document (or the part of the document that contains the dependencies we want to tackle) as one sentence. However, this approach fails, as the hypothesis recombination is inhibited, and the search space becomes much larger and computationally more expensive to handle (Hardmeier, 2014).

1.3.1.2 Document-level Decoder Docent

As we have exposed in section 1.3.1.1, the sentence-level approach does not allow us to model and integrate discourse level features into the decoder. Even though we tackle this issue in chapter 4 by passing information from previously translated sentences to the next ones, these kind of solutions are usually cumbersome and difficult to maintain.

Hardmeier et al. (2012) present a decoder called Docent[5] as part of the Disco-MT project (discourse-oriented Statistical Machine Translation),[6] which allows us to implement and integrate document-level features into the decoder. Instead of using the stack decoding algorithm described in section 1.3.1.1, Docent implements a search procedure based on local search, which works as follows. The decoder starts with an initial translation of the whole document, which is either randomly generated or obtained from a phrase-based SMT decoder. Next, at every stage of the search, the decoder randomly applies a state operation, such as `change-phrase-translation`, which replaces the translation of a phrase with another from the phrase table; `swap-phrases`, which exchanges phrases; `move-phrases`, which randomly moves phrases in the sentence; or `resegment`, which changes the segmentation of the source phrase. The decoder then computes the overall document score, taking into account the score obtained from each features function as in equation 1.1 and accepts a new state (i.e. a new translation of the document), if the document score of the current translation is higher than the last accepted.

In chapter 5, we describe a document-level feature that we integrated into Docent. This feature gives higher scores to document translations that contain semantically-similar translations of specific words in the text.

1.3.2 Neural Machine Translation

Phrase-based Statistical Machine Translation has shown to bring remarkable progress to Machine Translation over the last ten years. However, it is unnatural to split the input sentence into phrases and concatenate their translation. As a consequence, SMT does not even take into account the full sentence as a context, but only the direct surrounding context in each of the phrases. Neural Machine Translation emerged as a new MT paradigm in 2016, showing that it is able to better capture syntactic and semantic context, and achieving better performance than the state-of-the-art SMT systems in recent competitions (Luong and Manning, 2015, Sennrich et al., 2016a, Neubig, 2016).

1.3.2.1 Introduction to Artificial Neural Networks

Artificial neural networks are models that simulate how the neurons in the brain work. In contrast to machine learning algorithms, such as logistic regression, neural networks are able to learn complex non-linear hypotheses even with a large number of features.

[5] https://github.com/chardmeier/docent
[6] http://stp.lingfil.uu.se/~joerg/welcome.php?project=DiscoMT

To understand the parallelism between artificial neural networks and neural networks in the brain, we start with the most basic unit of such network: a neuron. A neuron is a brain cell that is designed to process and transmit information to a muscle or other cells through electrical pulses. Figure 1.2 illustrates how a neuron in the brain operates. In essence, the brain cell receives information from the dendrites, processes it, and transmits the output to other neurons through the axon. The axon terminals connect to the dendrites of other neurons constituting a neural network.

An artificial neural network is a group of interconnected nodes in a computer (i.e. artificial neurons) that operate like the network of neurons in the brain. Figure 1.3 illustrates a representation of such network. We observe that the network is constituted of multiple layers: (1) the input layer, which represents the input features; (2) the output layer, which outputs the final value; and (3) the hidden layer (or layers). Hidden layers are called hidden, since the values that the nodes in those layers use to make the corresponding computations come from the output of previous layers, and therefore, we do not see them in the training data

Neural networks were already used during the 1980s and early 1990s, but their use did not last longer, since training a system with neural networks is computationally very expensive compared to using other machine learning algorithms. However, computers evolved recently to be able to run large scale neural networks. Currently, artificial neural networks are the state-of-the-art technique for many machine learning applications in a wide range of areas such as biomedical, industrial, data mining, and financial.

The task of translating a text from one language to another is a machine learning problem, where the machine translation system aims at finding the translation hypothesis that maximises the model score. In SMT, we use linear models, which combine feature functions that model different aspects of translation quality such as the language model,

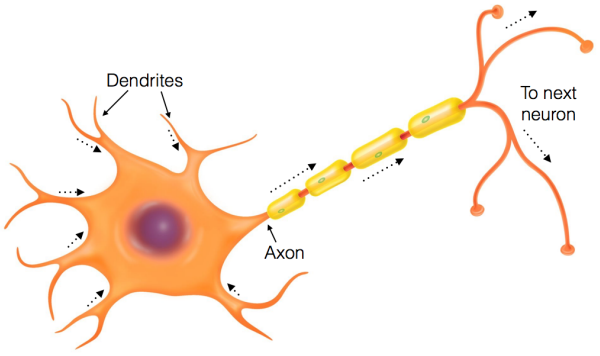


FIGURE 1.2: This image illustrates a neuron in the brain. The neuron gets input information from the dendrites and transmits the output to other neurons after processing the information through the Axon.

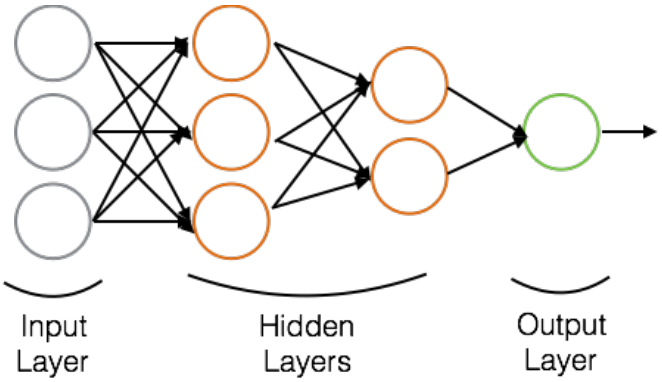


FIGURE 1.3: Representation of an artificial neural network with two hidden layers.

the translation model, and the reordering model to score a translation hypothesis, as described in section 1.3.1.1.

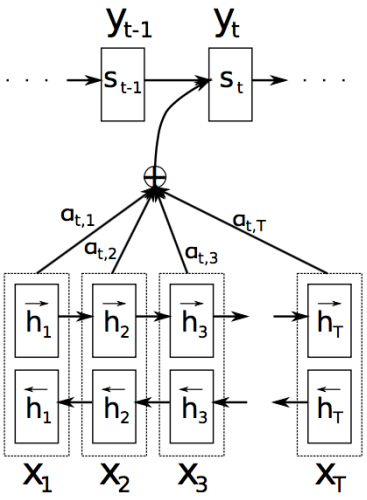
The issue with these linear models is that they cannot handle complex relationships between features. For example, linear models do not allow us to give a higher importance to the language model than to the phrase translation model, when the sentences are short and the opposite for long sentences. Neural networks can deal with these cases, outperforming state-of-the-art SMT systems in high-resource language pairs.

1.3.2.2 Attention-based Neural Machine Translation

The NMT systems that we build for the experiments follow the attention-based NMT architecture proposed by Bahdanau et al. (2015) and illustrated in Figure 1.4. In the following, we summarise the main features of this architecture, which consists of three components: the encoder, the decoder, and the attention model.

The encoder is a bidirectional recurrent neural network with gated recurrent units (Cho et al., 2014). Recurrent neural networks allow us to consider as context previously predicted words in an input sentence of any length. Specifically, given a sentence represented as a sequence of words $x = (x_1, \ldots, x_m)$, every word x_t gets as input the previous hidden state h_{t-1}, allowing the model to predict the current word in the context of the previously predicted words. Due to its sequential nature, recent words impact more on the prediction of w_t than those located further away. Therefore, we use gated recurrent units, which control how much of the information in the previous hidden state must be considered. This approach allows us to better model long-distance dependencies in the same sentence, as the attention model only keeps relevant information that is needed to predict next words in the sequence.

The bidirectional recurrent neural network is constituted by two recurrent neural networks that read the source sentence forwards and backwards, respectively, allowing the



Source: Bahdanau et al. (2015)

FIGURE 1.4: Illustration of the attention-based NMT architecture predicting the word y_t given an input sentence $(x_1, \ldots, x_T)$.

model to consider both the left and the right context at time t. Therefore, given an input sentence $x = (x_1, \ldots, x_m)$, the encoder calculates a sequence of annotation vectors $(h_1, \ldots, h_m)$, where each annotation vector h_i results from the concatenation of the hidden state $\overrightarrow{h}_i$ produced by the forward recurrent neural network and $\overleftarrow{h}_i$ from the reverse.

The decoder is a recurrent neural network that integrates an attention model, generating the target sentence $y = (y_1, \ldots, y_m)$, one word at a time. To predict a word y_i, the decoder considers a hidden state s_i, the previous predicted word y_{i-1}, and a context vector c_i to compute the attention scores. The context vector is based on the sequence of vector annotations $(h_1, \ldots, h_m)$ generated at the encoding stage and summarises the whole input sentence, giving different importance to each word. Specifically, c_i is the weighted sum of the annotations h_i, where each weight is computed through an alignment model. The alignment model is a neural network with a single hidden layer that gives the probability of y_i being aligned to x_i.

1.3.3 Neural versus Statistical Machine Translation

The successful performance of NMT systems, higher than state-of-the-art phrase-based SMT systems for several language pairs, has attracted the interest of the MT community, which suggests that NMT is a possible new paradigm in MT. Hence, several researchers have recently analysed and compared the output of Neural and Statistical MT systems to find out about their strengths and weaknesses. In these analyses, they mostly address the

English-German language pair, as it is specially challenging for MT due to the syntactic and morphological differences between these two languages.

Bentivogli et al. (2016) are the first to conduct an automatic analysis for this language pair on the translation of transcribed TED talks. To compare between Neural and Statistical MT, they evaluate the best four ranked systems at the IWSLT 2015 evaluation campaign (Cettolo et al., 2015),[7] a NMT system and three phrase-based SMT systems.

The evaluation results show that, in general, the NMT system outperformed all other systems in terms of overall translation quality. Indeed, the NMT system produces better morphological translations than the phrase-based systems. Specifically, it generates 19% less errors than the others in this category.

One of the most difficult challenges in SMT when dealing with the English-German language pair is word order, as it is very different between them. Specially, German has the peculiarity of placing verbs at the end of the sentence in some grammatical constructions, requiring long-range reordering, which is difficult to handle for phrase-based SMT systems. The analysis shows that NMT reduces the errors related to placement of verbs by 70% compared to the phrase-based systems, and word order errors in general by 50%.

Despite all these improvements over the phrase-based systems, the authors suggest that NMT needs to improve some other translation aspects. Long sentences, for example, are an issue in MT, and they are specially difficult for NMT systems. The results show that the translation quality of the NMT system degrades more rapidly for sentences longer than 35 words than in the other systems. Additionally, NMT also fails at handling cases that require a deeper understanding of the sentence semantics, such as the reordering of prepositional phrases and the detection of the focus of negation in the source sentence.

Popović (2017) extends the analysis by considering language-related issues that arise when German is the source language and English the target, such as mistranslated German compounds and the English continuous verb tenses, which do not exist in German. For the experiments, she analyses a NMT (Sennrich et al., 2016a) and a phrase-based SMT system (Williams et al., 2016) that are trained on WMT'16 news domain data.[8]

Popović (2017) found that NMT outperforms phrase-based SMT for German→English in dealing with morphology issues, such as German compounds and verb forms, and reordering. However, English continuous tenses are better handled by phrase-based SMT,

[7] http://workshop2015.iwslt.org
[8] Resulting annotated texts: https://github.com/m-popovic/german-english_pbmt-nmt-issues

which is the most frequent problem in the translation from German into English, followed by prepositions, which are an issue for both Neural and phrase-based SMT.

An important finding of Popović (2017)'s analysis is that the errors between the systems are complementary, since the majority of sentences present low error overlap between NMT and SMT. This indicates that MT could greatly benefit from a combination of both NMT and phrase-based SMT approaches.

Koehn and Knowles (2017) also report on the NMT challenges mostly for German-English and English-Spanish, but also for other language pairs, such as Czech-English, Romanian-English, and Russian-English. As Bentivogli et al. (2016), they find that NMT does not handle well long sentences.[9] Specifically, the authors show that the SMT system outperforms NMT on sentences longer than 60 words for English→Spanish.[10]

In addition, the authors find that "NMT systems have a steeper learning curve with respect to the amount of training data." That is, the quality of the NMT system is directly proportional to the amount of data, obtaining better performance for high-resource language pairs than for languages with less available data.

In the experiments, they also show that when the NMT system is tested on out-of-domain data, it still produces a fluent output, but at the cost of adequacy. For example, the German sentence *Schaue um dich herum* ("Look around you") is translated by an out-of-domain NMT system into *Take heed of your own souls.*

These results are in line with the findings of the WMT'16 competition Bojar et al. (2016). To evaluate the results of the submitted systems, they performed a manual evaluation of fluency and adequacy of their translation output. While the results showed that fluency is the main strength of NMT systems, there was no big improvement in terms of adequacy. The Neural MT systems submitted by Sennrich et al. (2016a) were the best constrained[11] system for 7 out of the 8 translation directions in the manual evaluation. These systems achieved an improvement of 13% points over the ONLINE-B system, an online statistical MT system. However, the improvement of adequacy was only about 1%. The fact that the output is fluent, but the meaning from the source sentence is not preserved is a real issue for manual post-editing or manual evaluation of the translation quality, as it increases the difficulty of identifying translation errors.

[9] Popović (2017) does not experiment with sentence length.
[10] Sentences longer than 60 words are very infrequent.
[11] The models were trained only on the provided data

1.3.4 Technical Settings

In our phrase-based SMT experiments, we use Moses (Koehn et al., 2007),[12] a framework to a automatically train translation models, to build our translation systems following the standard settings (Koehn et al., 2003). We use Giza++ (Och and Ney, 2003) to generate word alignments between each parallel sentence in the training corpora and to learn the corresponding phrase pairs. Additionally, we build a 5-gram language model using KenLM (Heafield, 2011). We finally tune our systems with Minimum Error Rate Training (Och, 2003b) on the corresponding development set, which is specified in each experiment.

In contrast, to build our NMT systems we use Nematus (Sennrich et al., 2017),[13] an open-source implementation for NMT, and the sample scripts and configuration files released by Sennrich et al. (2016a).[14] We then encode the words from our data via joint byte pair encoding (BPE) to enable open-vocabulary translation (Sennrich et al., 2016b).[15] The vocabulary size for all models is 90,000.

The total size of the embedding layer is 500 for the baseline and also for the systems trained with the lexical chains and word senses in chapter 6 and the dimension of the hidden layer is 1024. This allows us to fairly compare the systems, since a higher word dimension in the factored systems would improve their performance only due to the increase in the number of model parameters. We therefore distribute the total embedding size among the factors equally.

We always train the models for about a week, using Adam (Kingma and Ba, 2015) to update the model parameters on minibatches of size 80. Every 10,000 minibatches, we validate our model via BLEU and perplexity. In our NMT experiments, we use newstest2010 to tune our systems for both language directions German→French and German→English. The maximum length of the sentences is set to 50, and longer sentences are skipped.

1.3.5 Description of the Corpora used for the Experiments

In this section, we describe in the following the main corpora used in the development of the experiments for this thesis. The selection represents a broad range of genres and topics, such as news articles, movies subtitles, transcribed talks, proceedings of the

[12] http://www.statmt.org/moses
[13] https://github.com/rsennrich/nematus
[14] https://github.com/rsennrich/wmt16-scripts
[15] https://github.com/rsennrich/subword-nmt

European Parliament and essays on the alpine domain. The additional data used in some of the experiments is described in the corresponding chapter.

Europarl (v7) is a parallel corpus built from the proceedings of the European Parliament (Koehn, 2005).[16] The topics covered and the kind of language used are indeed narrowed in the context of the European Parliament. The corpus includes texts in 21 European languages that can be sentence-aligned in language pairs.

The WIT^3 corpus stands for Web Inventory of Transcribed and Translated Talks (Cettolo et al., 2012).[17] It is a collection of transcriptions and their translations in more than one hundred languages of the talks performed at the TED conferences.[18] In this corpus, we find transcribed speech on almost any topic: technology, entertainment, science, global issues, and so on.

As mentioned above, both Europarl and WIT^3 corpora come from the oral language. We also add in this category, the Open Subtitles corpus (Lison and Tiedemann, 2016),[19] a collection of translated subtitles from movies. In contrast, we also use the News Commentary (v11) (Tiedemann, 2012a)[20] and Text+Berg (Bubenhofer et al., 2013)[21] corpora, which are both extracted from written texts. The former are news on politics and economics. The latter is a collection of parallel German and French documents from the alpine domain, which was built as a result of digitising and processing the Swiss Alpine Club yearbooks from 1957 to 2016 (Volk et al., 2010).[22]

In addition, we often use in our experiments the WMT'16 test sets for tuning and testing.[23] These test sets are composed of news articles on different topics, written in a formal language and with a sentence length of about 30 words on average, which can be considered relatively long.

1.4 Relation to our own Published Work

Some of the experiments presented in this thesis have been previously published in different conferences. In the following, we chronologically list the conference papers by date

[16] http://www.statmt.org/europarl
[17] https://wit3.fbk.eu
[18] https://www.ted.com
[19] http://www.opensubtitles.org
[20] http://www.casmacat.eu/corpus/news-commentary.html
[21] http://textberg.ch/site/en/corpora/
[22] The entire Text+Berg corpus comprises so far all yearbooks from 1864 to 2016, although there were no parallel German-French documents until 1957.
[23] http://www.statmt.org/wmt16/translation-task.html

of publication and explain their relation to some sections in the thesis. In addition, we also describe the contribution of the other authors, when there was a clear separation of tasks.

- I presented the work on nominal references to German compounds described in section 4.2 at the bi-annual Konvens Conference on Natural Language Processing, which took place in Hildesheim (Germany) in October 2014 (Mascarell et al., 2014).
- We then extended the work on references to compounds by including Chinese-English and automatic post-editing in section 4.2.4. Xiao Pu presented the results at the Student Research Workshop of the International Joint Conference on Natural Language Processing (IJCNLP) and the Annual Meeting of the Association for Computational Linguistics (ACL) in Beijing (China) in 2015 (Pu et al., 2015). In this paper, Xiao Pu and Andrei Popescu-Belis focused on the Chinese→English experiments and the `Post-editing` approach, whereas I focused on the German→French experiments and the `Decode` method.
- We later published an overview of document-level Statistical Machine Translation at the MultiLingual magazine (Mascarell et al., 2016), a magazine that covers a wide variety of language-related topics.[24] I described the problem of incorrect translation choice of nouns in SMT. Annette Rios contributed to the final version of the article and described the issue of pronoun translation.
- Xiao Pu presented the experiments on consistent translation of repeated nouns detailed in section 4.3 in Valencia (Spain) at the Conference of the European Chapter of the Association for Computational Linguistics (EACL) in April, 2017 (Pu et al., 2017). Here, Xiao Pu and Andrei Popescu-Belis carried out the Chinese→English translation experiments and focused on the syntactic features. I focused mostly on the German→English translation task and the semantic features.
- Part of the work presented in chapter 5 was included in a paper that I presented at the Workshop on Discourse in Machine Translation held in Copenhagen (Denmark) in September 2017 (Mascarell, 2017).
- I presented the material explained in chapter 6 at the Conference on Machine Translation in Copenhagen (Denmark) in September 2017 (Rios et al., 2017). While Annette Rios and Rico Sennrich built the test set and focused on the evaluation of the systems, I developed the method to integrate lexical chains and sense embeddings in NMT.

[24] https://multilingual.com

1.5 Thesis Outline

In chapter 2, we give an overview of the state-of-the-art of discourse in Machine Translation, focusing on translation consistency and cohesion, which are the topics mostly covered in this thesis. We also give a background on the translation of pronouns and discourse connectives, since there is a large community working on these topics (specially on pronouns) and we carry out a discourse error analysis in chapter 3 where we also include them.

The error analysis presented in chapter 3 starts with a description of the guidelines followed to annotate the discourse-related translation errors in different types of data sets. These discourse errors are distributed into several categories concerning content words (i.e. nouns, adjectives, and verbs), pronouns, and discourse connectives. We then describe the annotation results and the main findings of the evaluation.

In chapter 4, we experiment with translation consistency in particular scenarios where a consistent translation is expected. In the first part of the chapter, we present a method that detects references to compounds, where the reference is the nominal head of the compound, and enforces the references to use the translation from the compound. We report results on the translation from German into French, and we then extended the approach to Chinese-English. In the second part, we tackle the consistent translation of pairs of repeated nouns for Chinese→English and German→English, using classifiers trained on syntactic and semantic features.

We then move on to lexical chains in chapter 5. Specifically, we benefit from context provided by the lexical chains of the source document to improve the translation. The proposed method detects first the lexical chains in the source using word embeddings and then keeps the semantic similarity of the words in the lexical chains in their counterpart target chains. For this purpose, we build and integrate a feature function into the discourse-oriented decoder Docent. We compare the performance of our method with an approach that uses a external lexical resource instead of word embeddings to detect the lexical chains. Additionally, we analyse the properties of lexical chains and their impact in translation.

In chapter 6, we continue with the work on lexical chains, integrating them into NMT. We use the method explained in chapter 5 to detect the lexical chains on the source side, and then include them as additional input factors in the data. We evaluate our method on a Word Senses Disambiguation task for German→English and German→French, which was especially designed to assess the performance of NMT systems.

Finally, chapter 7 concludes with a summary of the achieved results tackling the research questions listed in section 1.1.

Chapter 2

Background on Discourse in Machine Translation

This chapter gives the reader an overview of the research done in the literature related to our work. Our main focus of attention is improving lexical choice in Machine Translation, and therefore, we summarise the main approaches used for this purpose, such as applying topic modelling or cohesion models in section 2.1 and encouraging consistent translation in section 2.2. Additionally, even though we do not tackle the translation of discourse connectives and pronouns, we briefly describe their background in section 2.3 and section 2.4, respectively, since they have been actively addressed by researchers working on discourse, and we also consider them in a error annotation task described in chapter 3.

2.1 Cohesion in Translation

The main problem of sentence-level MT systems is that they deal with the sentences in a document independently. However, sentences function as a unit, defining document properties such as cohesion and coherence. While coherence has to do with the semantic meaningfulness of the text, cohesion concerns the connection between the sentences in the document.

One way to improve lexical cohesion and lexical choice is by topic modelling, using methods based on Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA). Zhao and Xing (2006)'s approach improves the word-alignment by applying bilingual topic models, which leads to gains in translation quality, and Tam et al. (2007) propose a bilingual LSA adaptation of the language model and translation lexicon. While these

works use a bilingual LDA, Gong et al. (2010) use a monolingual LDA to introduce document topic in translations from Chinese to English. Later, Hasler et al. (2014) exploits a bilingual LDA model to obtain the phrase translation probabilities used during training from the phrase table.

Some other studies focus on lexical cohesion to assess the quality of the translation at document-level. Wong and Kit (2012) integrate lexical cohesion devices (i.e. semantically related words) in automatic evaluation metrics, such as repetition, synonyms or near-synonyms, and hyponyms or hypernyms. The authors suggest that the quality of the translation is directly proportional to the amount of cohesion devices used.

Xiong et al. (2013a) are the first to successfully integrate lexical cohesion into discourse-oriented MT. They focus on lexical cohesion devices and develop models to capture and encourage lexical cohesion in the translation by rewarding the occurrence of cohesion devices. The models are integrated into a hierarchical phrase-based MT system trained and tested on Chinese-English parallel data. The authors show a significant improvement over the baseline and conclude that, contrary to the findings reported by Wong and Kit (2012), lexical cohesion devices should be used appropriately rather than frequently. Later, Lapshinova-Koltunski (2015) finds that their use depends on the genre and language.

Gong et al. (2015) develop document-level evaluation metrics based on topic modelling and simplified lexical chains, which consider only repeated words in the text (Gong and Zhou, 2015), and integrate them into existing traditional evaluation metrics. Their approach achieves a better correlation with human judgments on the evaluation of MT quality than traditional metrics.

Following Xiong et al. (2013a)'s work on lexical cohesion, Xiong et al. (2013b) present a method that uses lexical chains as a means to keep the document cohesion in the translation and improve the lexical choice of the words in the chain. They use a external lexical resource to detect the lexical chains in the source Chinese document, and they then create the counterpart lexical chains in the English translation using maximum entropy classifiers. These classifiers are trained on each word in the source chain and predict the translation of each word based on the previous and the next word in the chain and the immediate surrounding context. The resulting lexical chains are used by their cohesion models, which are integrated into a hierarchical phrase-based MT system Chiang (2005), reporting a substantial improvement of the system's lexical choice for Chinese→English.

Since Xiong et al. (2013b)'s classifiers are trained on each word in the lexical chains extracted from the training data, the approach presents a drawback for words from the

lexical chains detected in the test set that are infrequent or missing in the training data. In chapter 5, we present a method to integrate lexical chains in the document-oriented decoder Docent that overcomes this issue and evaluate its performance on the German→English translation task. Additionally, we extend our method to integrate document-level knowledge from lexical chains in Neural MT and assess the performance of the method on a Word Sense Disambiguation task in chapter 6.

Later, Xiong and Zhang (2014) integrate a sense-based translation model into SMT that takes advantage of word senses induced from the surrounding contexts in which the word occurs. The model is based on maximum entropy classifiers that use the senses to predict the best translation of each word within a context window consisting of the previous and next n-words. In chapter 6 we train a system with word senses to evaluate whether Neural MT can benefit from this knowledge to improve lexical choice.

Similarly to the work on lexical chains, Xiong and Zhang (2013) present a topic-based coherence model that, given a source coherence chain, predicts its target coherence chain, encouraging the decoder to make coherent lexical choices. The experimental results show a marginal improvement over the baseline for Chinese→English.

Some researchers also focus on the evaluation of coherence, which is often performed on shuffled coherent data. Smith et al. (2016) tackle coherence as part of the MODIST project (i.e. modelling discourse in translations) and take a step forward automatically evaluating translated documents instead.[1] This results in a more challenging task as the translated text do not have the artificial shifts of focus that result from shuffling. They propose a syntax-based coherence model that is able to correctly score human translations higher than MT output, outperforming other existing models.

2.2 Encouraging Lexical Consistency

There have been several attempts to improve the lexical choice of MT by encouraging the systems to consistently use the translations throughout the document. Lexical consistency (i.e. repetition of the same translation) is related to cohesion, since repetition is a referential device that together with other cohesion devices, such as ellipsis, substitution, lexical cohesion and conjunction, makes a document cohesive.

The idea stems from the one-sense-per-discourse hypothesis by Gale et al. (1992), which states that "well-written discourses tend to avoid multiple senses of a polysemous word."

[1] http://staffwww.dcs.shef.ac.uk/people/L.Specia/projects/modist.html

Hence, for example, the English polysemic word *bat* only appears in one of its senses (i.e. as the flying animal or the paddle) in a document.

These findings are later investigated and applied in the context of MT to assess whether words that have multiple translations in the target language should be consistently translated (i.e. using the same translation) across the document (Carpuat, 2009). The experiments are carried out for French-English on news articles, and the evaluation revealed that SMT systems translate quite consistently because of their low variability in lexical choice. However, when the hypothesis does not hold, and she finds more than one translation, it is often due to wrong lexical choices. Furthermore, Carpuat (2009) reports improvements on the translation quality when the discourse hypothesis is enforced.

Some research in SMT focused later on analysing and encouraging consistency in translation, based on (Carpuat, 2009)'s one-translation-per-discourse constraint. Carpuat and Simard (2012) conduct later an in-depth analysis of consistency in the SMT output compared to human translations. They experiment with several English-French and Chinese-English phrase-based SMT systems trained on different conditions, such as data size and genre. The study concludes that SMT is already fairly consistent at lexical choice, nearly as consistent as human translation. Furthermore, consistency is not related to translation quality. Indeed, higher consistency levels are achieved with weaker systems (i.e., trained on less data), since they have less vocabulary choices. Nevertheless, incorrect lexical choices are usually attributable to inconsistent translations, so inconsistency cannot be ignored and needs to be tackled.

Tiedemann (2010) proposes a cache-model to enforce consistent translation of phrases across the document. However, the problem with the use of a cache is that it easily gets contaminated. That is, bad translations can be stored in the cache, and those translation errors propagate to the following sentences in the document. Gong et al. (2011) extend Tiedemann (2010)'s approach using a dynamic, static, and topic caches to store document-level information, reporting an improvement in the translation quality for Chinese-English. Similarly to Tiedemann (2010)'s cache, the dynamic cache proposed by Gong et al. (2011) stores the bilingual phrase pairs from the previously translated sentences. To mitigate the error propagation issue, they use a static cache initialised with phrase pairs from similar documents at the beginning of the translation. Finally, their approach keeps the noisy translation from the previous two caches in check with the use of the topic cache, which stores relevant topic words, as a means to apply topic modelling.

Xiao et al. (2011) and later Martínez Garcia et al. (2014) propose a method to deal with inconsistencies at post-processing stage for English-Chinese and English-Spanish, respectively. Specifically, they first identify the ambiguous words in a document and

store their translations. Next, they modify the identified words in the translation output with their consistent translation. Martínez Garcia et al. (2014)'s approach differs from Xiao et al. (2011) as the latter identifies the ambiguous words and their translation after decoding.

Ture et al. (2012) analyse the one-translation-per-discourse constraint on newswire documents translated from Arabic into English using a hierarchical phrase-based MT system (Chiang, 2005). They observe that in 128 of the 176 cases where the system offers multiple translation choices, a human makes a consistent choice, and the remaining cases are often result of stylistic choices. Ture et al. (2012) suggest that instead of imposing consistency across the document, we should further study when and how to apply it, as language-specific phenomena may require linguistic variation in some cases. They then integrate a set of cross-sentence consistency features to the translation model, reporting substantial improvements for Arabic-English and Chinese-English.

Guillou (2013) tackles consistency in a different way, analysing *where* (i.e. part-of-speech) lexical consistency is desirable for English-French in several text genre, such as news articles, novels, natural science texts, instruction manuals, and public information. The results suggest that nouns should be encouraged to be translated consistently throughout the document, across all genres. Additionally, consistent translation of rare verbs and adjectives is beneficial for technical reports, whereas only adjectives for public information documents. Guillou (2013) concludes in line with Carpuat (2009) and Carpuat and Simard (2012) that consistency is high on average and that inconsistencies in SMT often lead to incorrect lexical choices. However, consistency is not always desirable and should be selectively enforced. For example, low frequent verbs, defined as light verbs, are inconsistently translated by human translators and that should be reflected in the SMT output.

Hardmeier et al. (2012) integrate a feature in the document-oriented decoder Docent that encourages the use of semantically similar words in the translation, using a 30-dimensional word space model based on Latent Semantic Analysis. Specifically, the decoder uses the model to assess the adequacy of the translation of content words in their preceding context. Despite reporting small gains in translation quality for English-French, Hardmeier et al. (2012) claim that cross-sentence models should be further examined. Martínez Garcia et al. (2015) present later a similar approach that computes the adequacy of translated words using monolingual and bilingual embedding models for English-Spanish. They also report slight improvements with the bilingual model.

Zhang and Ittycheriah (2015) develop document-level features for the source and target to improve lexical choice and consistency in translation. The authors report quality

improvements in the translation of newswire and weblog documents from an Arabic dialect into English.

More recently, Martínez Garcia et al. (2017) implements a feature for the document-level decoder Docent that uses word embeddings to translate repeated words consistently, and evaluates its performance on the English-Spanish translation task. Word embeddings are representations of words in a vector space, which proved to provide good performance at computing the similarity between words even across languages (Mikolov et al., 2013). The manual evaluation reveals that 60% of the time the output improves over the baseline and 20% of the time is equivalent or equal.

As for NMT, Wang et al. (2017) propose a method to consider a wider context from source-side previous sentences, which, together with Jean et al. (2017), is the first attempt to integrate discourse into NMT. The method summarises the discourse context using a hierarchy of Recurrent Neural Networks, reporting a substantial improvement in automatic evaluation scores for Chinese→English. Specifically, the authors perform a small manual evaluation on 15 randomly selected documents and observe that their approach fixes 76% and 75% of the ambiguity and consistency errors, respectively. Our method presented in chapter 6 is an independent early attempt to integrate discourse in NMT, as the work of Wang et al. (2017) and Jean et al. (2017).

Following Guillou (2013)'s findings, we encourage consistency in specific scenarios where a consistent translation is expected. In addition, we focus mostly on nouns, as their consistent translation is more desirable than other parts-of-speech (Guillou, 2013). In section 4.2, we attempt to improve the lexical choice of nouns that refer back to nominal compounds by using the translation of the compound. We also address consistency in the translation of pairs of repeated nouns in a document, using a machine learning classifier based on syntactic and semantic features (see section 4.3).

2.3 Discourse Connectives

Discourse connectives are words or phrases that signal a discourse relation between coherent sentences in a document, contributing to the understanding of the document. The translation of connectives poses a challenge not only for MT systems, but also for human translators.

Cartoni et al. (2011b) analyse the variability of discourse connectives in French texts from the proceedings of the European Parliament, which are originally French or translated from English, German, Italian or Spanish. They find that discourse connectives show more variability than other lexical items, as human translators decide to express them

implicitly (i.e. the connective is not translated) or explicitly depending on the source language. In line with Cartoni et al. (2011b)'s work, Hoek et al. (2015) report that the implicitness or explicitness in the translation of a connective depends on the language pair and the expectedness of the relation, and that human translators vary more the translation of connectives than MT.

Ambiguous discourse connectives are specially difficult to translate for MT systems when they have several translations in the target language. For example, the English connective *since* can indicate either a temporal or a causal discourse relation.

Meyer et al. (2011) and Meyer (2011) manually annotate the senses of discourse connectives (e.g. contrast, temporal, or causal) to train classifiers that are able to automatically predict their senses. Based on the obtained annotations, Meyer (2011) and Meyer and Popescu-Belis (2012) modify the phrase-table of the English→French SMT decoder to encourage correct translations for specific senses. Additionally, Meyer and Popescu-Belis (2012) train the SMT system on labeled data (i.e. connectives are labeled with their sense), so the system directly learns the correct translation for each sense. The experimental results show that both approaches achieve an improvement in the translation of English discourse connectives. Similarly, Meyer and Poláková (2013) report improvements on the translation of connectives from English into Czech, using also a system trained on labeled data. Finally, Meyer et al. (2012) report quality gains by training a phrase-based English→French SMT systems indicating the connective senses as an additional factor (e.g. *while*|*contrast*) in combination with part-of-speech tags.

Cartoni et al. (2011b)'s corpus study and all Meyer's work (Meyer et al., 2011, Meyer, 2011, Meyer and Popescu-Belis, 2012, Meyer and Poláková, 2013, Meyer et al., 2012) belong to the research done in the COMTIS Sinergia project.

2.4 Pronominal Anaphora and Pronoun Translation

One way to create a cohesive text is through the use of references, which refer to an entity that appears forward or backward in the discourse. Pronominal anaphora is defined as the use of (anaphoric) pronouns to refer to an entity that appears earlier in the document and, together with pronoun translation, is a very challenging issue for MT (Le Nagard and Koehn, 2010, Hardmeier and Federico, 2010).

Some anaphoric pronouns are well translated without knowing about the antecedent. However, some linguistic information, such as the gender of the antecedent, may change from the source to the target language, and therefore, the gender of the pronoun as well. In such cases, we need the information from the antecedent provided by the coreference

resolution systems. However, the performance of coreference resolution still leaves room for substantial improvements. Other problems are related to the use of pro-drop languages in the source, such as Spanish or Italian, since they tend to elide subjects and MT systems need to produce the right pronoun in the translation (Rios Gonzales and Tuggener, 2017).

Some researchers focus on the development of tools to improve anaphora resolution. For example, Tuggener (2016) implements CorZu, an incremental coreference resolution system for German.[2] This system was later adapted to the pro-drop Spanish language, and its performance was evaluated on the SemEval 2010 competition data set,[3] achieving better accuracy for elided subjects than the best performing system of the task (Rios Gonzales and Tuggener, 2017). In contrast, Hardmeier et al. (2013) address the translation of third-person subject pronouns from English into French and propose a pronoun prediction method modelled in a neural network architecture that does not require annotated data. Also, Novák and Žabokrtský (2014) present a system for Czech-English.

To support with the development of better anaphora resolution systems, researchers have manually annotated coreferences on different types of corpora. Guillou et al. (2014) release a corpus with manual pronoun-coreference annotations from a collection of English-German and English-French documents.[4] In a test set from the proceedings of the European Parliament, Popescu-Belis et al. (2012) label the English pronouns with their corresponding counterpart in the French translation, without annotating the antecedent.

Translation of pronouns has been a focus of attention in a recent series of cross-lingual pronoun prediction competitions (Hardmeier et al., 2015, Guillou et al., 2016, Loáiciga et al., 2017), which cover several language pairs such as English-German, English-French, and even Spanish-English in its last edition. The task of these competitions consist of a parallel source-target text, where the pronouns that need to be predicted are marked on the target side. This way, other translation errors do not interfere with the translation of the pronouns and the task is fully focused only on the pronoun prediction. As a result of these competitions, Guillou and Hardmeier (2016) released a test suite to automatically evaluate the translation of pronouns.

Recently, Jean et al. (2017) proposed an extension of the state-of-the-art attention-based NMT architecture (Bahdanau et al., 2015) that takes into account the context from surrounding sentences. They evaluate the performance of the models on the WMT'16

[2] https://github.com/dtuggener/CorZu
[3] http://stel.ub.edu/semeval2010-coref/
[4] http://opus.lingfil.uu.se/ParCor

cross-lingual pronoun prediction task (Guillou et al., 2016), reporting benefits from a wider-context when the models were trained on small corpora, but not with a larger corpus.

2.5 Summary

There has been a number of studies to tackle discourse-related issues in Machine Translation. In this chapter, we summarised the relevant literature concerning the translation of content words, such as verbs and nouns, pronouns, and discourse connectives. The latter was successfully addressed in the COMTIS Sinergia project (Meyer et al., 2011, Meyer, 2011, Meyer and Popescu-Belis, 2012, Meyer and Poláková, 2013, Meyer et al., 2012) and followed in a more linguistic-oriented manner in the MODERN project by Hoek et al. (2015). The translation of pronouns, which is still a very challenging issue for MT, has recently gained substantial attention with the introduction of cross-lingual pronoun prediction competitions (Hardmeier et al., 2015, Guillou et al., 2016, Loáiciga et al., 2017). Both the translation of pronouns and connectives are not the focus of this thesis, but we consider them relevant for comparison in a discourse error analysis that we address in the next chapter.

To improve the translation of content words we distinguish between a line of research that encourages translation consistency (Carpuat, 2009, Carpuat and Simard, 2012, Guillou, 2013, Martínez Garcia et al., 2017, Wang et al., 2017) that we address in chapter 4 and another focused on the development of cohesion models (Xiong et al., 2013a,b, Hasler et al., 2014) related to chapter 5 and chapter 6. As reported by Hardmeier (2014), we observe in their work that the improvements in translation quality are usually marginal except for the Chinese-English language pair (Xiong et al., 2013b, Wang et al., 2017). This shows that it is easier to obtain greater gains with a semantically dissimilar language pair, since it is more challenging for the baseline system to generate good translations without the help of additional model features (Hardmeier, 2014).

Chapter 3

Analysis of Machine Translation Errors

In this chapter, we define and perform an annotation task of discourse errors in translation, using an out-of-domain phrase-based SMT system to translate from German into French, which are the most spoken official languages in Switzerland, and from English into Spanish, as they are a high-resource language pair widely used in research. The goal of this annotation task is twofold: to detect discourse errors in the translation and assess whether these errors can be tackled using discourse knowledge.

Specifically, we annotate incorrect translations of connectives, pronouns and content words (i.e. common nouns, adjectives, and verbs) on different text genres, such as news articles, transcribed and translated talks, and subtitles. Additionally, we also annotate whether local or discourse context would help to improve the lexical choice of content words (section 3.1). We finally discuss the annotation results in section 3.3.

3.1 Annotation of Discourse Translation Errors

This section describes our guidelines to annotate translation errors related to discourse. The annotation is performed by two different annotators. While one tackles the discourse translation errors for German→French, the other focuses on English→Spanish.

In the annotation task, we focus on content words, pronouns, and connectives. Specifically, the content words we take into account are common nouns, adjectives, and verbs. We ignore mistranslations of proper nouns, since they can be avoided using a named entity tagger (e.g. the Swiss ice hockey coach *Patrick Fischer* could be mistranslated into *Patrick fisherman*). Furthermore, we only consider non-separable verbs in German,

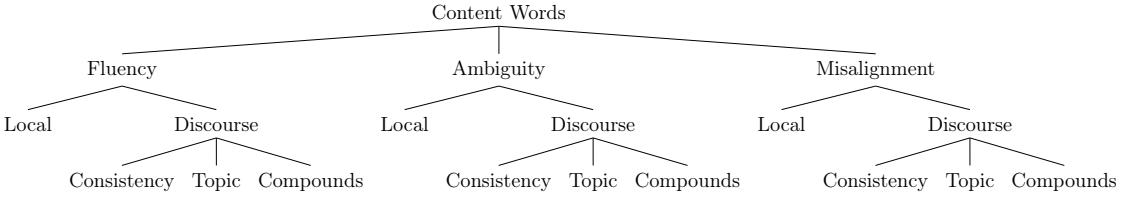


FIGURE 3.1: Diagram of the translation error annotation of content words: common nouns, adjectives, and verbs.

since this issue can be solved at sentence-level using syntactic information (Sennrich and Haddow, 2015).

The translation of pronouns and discourse connectives are not the main focus of this thesis. However, we include them in this analysis to get a better comparison of the main discourse errors found in the output of Machine Translation.

3.1.1 Annotation of Content Words: Nouns, Adjectives, and Verbs

We distinguish three types of discourse errors when dealing with content words: *fluency*, *ambiguity*, and *misalignment* errors. Figure 3.1 gives an overview of the annotation schema defined in this section.

While ambiguity errors occur when the Machine Translation system does not choose the right sense of the word, we find fluency errors when the translated word is not the preferable one in the document genre or there is a better translation in the context. The main difference between ambiguity and fluency errors is that in the latter both incorrect and correct translations belong to the same sense of the word (e.g. *persons* with disabilities versus *people* with disabilities). In the following, we exemplify an ambiguity error in a translation from German into French:

(3.1) **Source:** Im Februar 2010 lassen sich die drei Jugendfreunde Pascal Burnand, Gabriel Chevalier und Raphaël Houlmann, die schon von Kindsbeinen an mit den Gipfeln des Juras auf Du und Du sind, vom ***Bericht*** der fröhlichen Bande von damals inspirieren, die Route zu wiederholen ...
Arktische Atmosphäre Gleiche Fortbewegungsart, gleicher Enthusiasmus - der ***Bericht*** der Alten ähnelt demjenigen der Jungen, obschon 30 Jahre dazwischen liegen.

Machine Translation: En février 2010, les trois Jugendfreunde Pascal Burnand, Babriel Chevalier et Raphaël Houlmann, déjà de Kindsbeinen à avec les sommets du jura et sont, du ***récit*** de la joyeuse bande de l'époque, inspirent de

répéter la voie ...
De la côte de l' atmosphère, même moyen, même enthousiasme - le ***rapport*** de la vieille, et celui des jeunes, bien que trente ans entre.

Human Reference: En février 2010, trois copains d'enfance, Pascal Burnand, Gabriel Chevalier et Raphaël Houlmann, bercés par les cimes jurassiennes depuis leur plus jeune âge, s'inspirent du ***récit*** de la joyeuse clique des «anciens» pour refaire la route ...
Ambiances arctiques même moyen de locomotion, même enthousiasme des équipes, le ***récit*** des «anciens» ressemble à celui des plus jeunes, malgré la trentaine d' années écoulées .

Here, we observe that the German noun *Berichte* appears twice in the source document. We find both in the reference correctly translated into *récit* ("story"), whereas the SMT system incorrectly translates the second occurrence into *rapport* ("report").

In the third category, we find misalignment errors, which are those errors that do not belong to any of the other two categories and the translation does not correspond to any of the meanings of the source word. These type of errors may be due to wrong alignments of the training data used to build the system, as in example 3.2, where the English *modern* is mistranslated into *economía moderna* ("modern economy").

(3.2) **Source:** and during this time, there's a surge of prolactin, the likes of which a *modern* day never sees.
SMT: y durante ese tiempo, hay una oleada de prolactin del día nunca vio que una *economía moderna*.

Once we identify the type of discourse error, we annotate it according to how it could be solved. We then distinguish two different categories that differ from the amount of context that they need to take into account: *discourse* and *local* context. Local context considers only the surrounding words. Specifically, we limit the local context to three words to the left and three to the right. Anything that goes beyond three-words distance or crosses sentence boundaries falls into the discourse category. These categories are not exclusive. That is, some translations can be improved with both discourse and local context.

Incorrect translations that can benefit from discourse-level context must be annotated as *consistency*, *topic*, or *compounds*, which refer to the way discourse helps. If the mistranslated word appears repeatedly in the document, and the other occurrences are correctly translated, we can profit from consistency. That is, we could fix the errors by

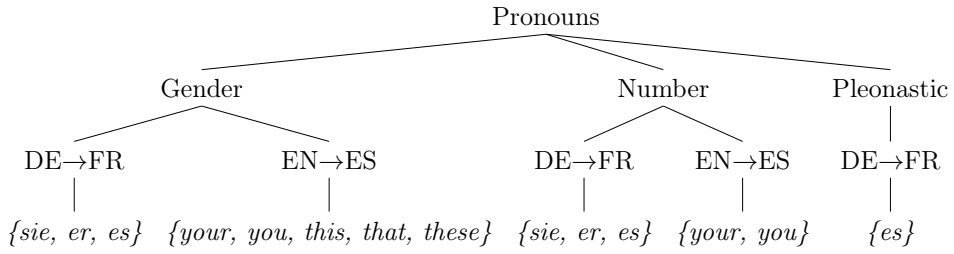


FIGURE 3.2: Diagram of the translation error annotation of pronouns.

using the same correct translation throughout the document. Deciding which is the right translation is also an issue, but we do not address it in the annotation task. In chapter 4, we study translation consistency and propose methods to address it.

In other cases, the mistranslated word does not fit in the topic of the context. We can find a better translation by looking at semantically-related words from the context, and therefore we label them as topic. For example, the English word *calf* cannot be translated into the Spanish *pantorrilla* (i.e. the fleshy lower leg) in a text about whales. The consistency category prevails over the topic one. That is, if the same word (e.g. *calf*) appears well translated in the surrounding sentences, we annotate it only as consistency. In chapter 5, we run several experiments that utilise the semantic-similar words in a document to improve lexical choice in SMT. We then extend the work to NMT in chapter 6.

The last category, compounds, is related to consistency. A noun belongs to this category if it refers back to a compound in the document, and it is the nominal head of the compound it refers to. For example, the German word *Typen* in the phrase *diese Typen* can be translated into English either as *types* or *guys*. However, knowing that it refers to the compound *Körpertypen* ("body types") in a previous sentence helps to disambiguate the word and translate it into *types*. We define this category, because German is a language rich in compounds, and therefore we expect some translation errors to fall in this category. We study consistent translation of references to compounds in section 4.2.

3.1.2 Annotation of Anaphoric Pronouns

We also want to analyse how often pronouns are mistranslated compared to content words. We focus on the translation of the German pronouns *sie*, *er*, and *es*, since they present some translation challenges related to gender and number, and the English *your*, *you*, *this*, *that*, *these*. Figure 3.2 shows the annotation schema of pronouns.

The German pronoun *sie* represents the third person singular and plural, and the formal form of *you*. Note that the formal form must be capitalised (i.e. *Sie*), but for the sake of simplicity, in these experiments we lowercase all data, and we do not keep this

distinction.[1] This pronoun is highly ambiguous as it can be translated into four different pronouns in French: *elle* (feminine singular), *elles* (feminine plural), *ils* (masculine plural), and *vous* (formal form).

Even if the system translates correctly the pronoun in gender and number, the translation of the pronoun might still be wrong. That is because the gender of pronouns is determined by their antecedent, which can change the gender in the target language. Example 3.3 illustrates this issue, when translating from German into French.

(3.3) **Source**: Das Maultier wird von einer wütenden Katze gejagt. *Es* hat Angst.
SMT: La mule est poursuivi par un chat en colère. *Il* a peur.
Reference: La mule est poursuivi par un chat en colère. *Elle* a peur.

In this example, we see that the gender of *mule* is neutral in German (*das Maultier*) and feminine in French (*la mule*). The pronoun that refers to it should therefore change accordingly from *es* to *elle*. However, the system is not aware of the antecedent and translates *es* to the most likely translation *il*. The pronoun is not only incorrect, but it also changes the meaning of the sentence: now the cat is afraid. We annotate all mistranslations of *sie*, *es*, and *er* with the reason of the error: *gender*, *number*, or both.

Furthermore, the pronoun *es* is also pleonastic, when it does not have any antecedent. For example, *it* (pronoun *es* in German) is a pleonastic pronoun in the sentences *it* is raining, *it* seems that, or *it* was known. When this type of pronoun is mistranslated, we annotate it only as *pleonastic*.

Translating pronouns is not as problematic from English into Spanish as from German into French. For English→Spanish, we focus on the possessive pronoun *your*, the personal pronoun *you*, and the demonstrative pronouns *this*, *that*, and *these*. The problem with *your* and *you* is that they have a different translation in Spanish depending on the formality and the number of the antecedent: *tu* (*you* and *your*, both 2^{nd} person singular and informal), *tus* (*your* informal and plural), *su* (*your* formal and singular), *sus* (*your* formal and plural), *vosotros* (*you* 2^{nd} person plural and informal), *usted* (*you* 2^{nd} person singular and formal) or *ustedes* (*you* 2^{nd} person plural and formal).

The demonstrative pronouns in English do not make any distinction between gender, but Spanish does. Therefore, the system would need to know the gender of the antecedent to provide the correct translation among: *eso* (*that* masculine), *esa* (*that* feminine), *esto* (*this* masculine), *esta* (*this* feminine), *estos* (*these* masculine), *estas* (*these* feminine).

[1] It is possible to keep the formal *Sie* capitalised by truecasing the training and testing data.

3.1.3 Annotation of Ambiguous Discourse Connectives

We annotate only those connectives that are mistranslated as a result of an ambiguity problem. Consider the example 3.4 that translates from English into French from Meyer et al. (2015). In this example, the connective *since* gets translated into *parce que* (causal) instead of *depuis* (temporal).

(3.4) **Source**: What stands between them and a verdict is this doctrine that has been criticized *since* it was first issued.
SMT: Ce qui se situe entre eux et un verdict est cette doctrine qui a été critiqué *parce* qu'il a d'abord été publié.
Reference: Seule cette doctrine critiqué *depuis* son introduction se trouve entre eux et un verdict.

Table 3.1 lists the German pronouns that we tackle in the annotation task, and their respective French translations. As we mentioned, we only consider the pronouns that have multiple translations in French that are often non-interchangeable.

TABLE 3.1: German connectives taken into account for the annotation task and their translations into French.

German	**French**			
somit	ainsi	donc		
sowie	comme	ainsi que		
wo	alors que	puisque		
wenn	lorsque	quand	si	
wie	que	comme	ainsi que	
beziehungsweise	plus précisément	ou	respectivement	
als	comme	lorsque	quand	en tant que

TABLE 3.2: English connectives taken into account for the annotation task and their Spanish translations.

English	Spanish	English Example
but	pero (yet)	I may be young, but I am not naive.
	excepto (except)	She could not do anything but wait.
since	desde (temporal)	I live in Switzerland since 2013.
	porque (because)	I will eat your pizza, since you are on diet.
whether	si (if)	I am not sure whether to hike tomorrow.
	sin importar si (even if)	I will hike tomorrow whether or not it rains.
so	tan (very)	The views were so amazing...
	así que (therefore)	I forgot my keys, so I had to drive back.
	también (also)	I want to go to the beach and so does she.

In the translation from English to Spanish, we focus on the following English connectives: *since*, *but*, *whether*, and *so*. Table 3.2 shows the Spanish translations of each connective, together with an English example.

3.2 Setup of the Spanish→English and German→French SMT systems

We build German→French and English→Spanish out-of-domain systems that emulate general purpose SMT systems. That is, their purpose is to translate different type of texts. They are trained on data from Europarl, News Commentary, WIT^3, and Open Subtitles, which represent a mix of text types and genres. Additionally, the German→French system uses the Text+Berg corpus. See section 1.3.5 for a more detailed explanation of the data.

In order to train out-of-domain systems that are not biased towards a specific genre, we use the same amount of data from each corpus. Specifically, we pick from each corpus 2M tokens for training, 5K for tuning, and 2.5M for the language model. Table 3.3 details the corresponding number of sentences per language pair, and the total of data used. The systems are both trained using the standard settings (Koehn et al., 2003), and tuned with Minimum Error Rate Training (Och, 2003b) on a development set. We also build a 5-gram language model using KenLM (Heafield, 2011).

We build parallel test sets from the described data, where each of them belongs to a different corpus. Since Text+Berg is only used for the German→French translation task, we have four test sets in English→Spanish and five in German→French. The Spanish and French counterpart are kept as a reference during the annotation task and for evaluation purposes.

Each test set has a total of 300 sentences, but different number of tokens, as the average sentence length varies among the corpora. Table 3.4 shows the total amount of tokens of

Table 3.3: Data sets used for the annotation task.

	Training		Tuning		Language Model	
	en-es	de-fr	en-es	de-fr	en-es	de-fr
Europarl	79K	87K	205	220	94.5K	94.5K
Wit^3	125K	125K	290	310	158K	139K
News Commentary	89K	88K	250	215	95K	96.5K
Open Subtitles	310K	314K	980	950	445K	325K
Text+Berg	–	104K	–	340	–	130K
Total	603K	718K	1,725	2,035	792.5K	785K

TABLE 3.4: Total number of tokens per language pair of each test set.

	en-es	de-fr
EUROPARL	∼ 7.3K	∼ 7.3K
WIT^3	∼ 4.7k	∼ 7.3K
NEWS COMMENTARY	∼ 5.3K	∼ 5.4K
OPEN SUBTITLES	∼ 1.3k	∼ 1.5K
TEXT+BERG	–	∼ 4.6K

TABLE 3.5: BLEU scores of the test sets.

	en-es	de-fr
EUROPARL	31.71	25.48
WIT^3	27.35	18.46
NEWS COMMENTARY	25.24	22.86
OPEN SUBTITLES	10.49	13.35
TEXT+BERG	–	12.72

each test set. We observe that Europarl contains long sentences as it is the test set with the highest number of tokens. On the contrary, Open Subtitles contains short sentences.

Table 3.5 reveals the obtained BLEU scores on the test sets translated with the out-of-domain systems. We see that the lowest BLEU scores are obtained when translating the test sets from Open Subtitles and Text+Berg, because of different reasons. Text+Berg is hard to translate for our systems as it contains very specific vocabulary in the alpine domain. This type of text requires an in-domain system (i.e. a system trained only with Text+Berg data) to obtain better translations. As for Open Subtitles, we observe in a manual analysis of the test set that the reference translation is not a direct translation of the source text. The n-gram overlap between our systems' translation and the reference is then low, negatively affecting the BLEU scores.

Example 3.5 lists the first nine sentences of the source and the corresponding reference translation of the Open Subtitles test set, and we added a direct translation to compare it to the reference. The sentences that differ in meaning between the source and the reference translation are highlighted in italics. For example, the reference translation *Te llamaré a la oficina para ver a qué hora llegas* ("I'll call the office to check at what time you arrive") has a different meaning than the original source sentence *I'll call in 30 minutes to check.*

(3.5) **Source**

Go straight to the office...

Don't dawdle on the way

Don't worry

I'll call in 30 minutes to check
Oh, hello fujio
Is your mother here, too?
Why are you outside?
It's no fun listening to women's talk

Direct Human Spanish Translation
Ve directamente a la oficina...
No pierdas el tiempo por el camino.
No te preocupes
Voy a llamar en 30 minutos para comprobar
Oh, hola fujio
¿Está tu madre aquí también?
¿Por qué estás fuera?
No es divertido escuchar conversaciones de mujeres

Human Reference
Bien, entonces, adelante, a la oficina...
No pierdas el tiempo por el camino.
No te preocupes.
Te llamaré a la oficina para ver a qué hora llegas.
Oh, hola fujio .
- *Si*
¿Por qué estás fuera?
No es divertido entrar ahí con mi madre.

3.3 Analysis of the Annotated Discourse Errors

We analyse and summarise the results of the annotation errors in each test set in table 3.6. To get a fine-grained insight into the errors, the annotation of content words is subdivided into common nouns, adjectives, and verbs for each of the error categories: ambiguity, fluency, and misalignment. The error percentage is computed as the percentage of the errors to the total of possible cases. For example, the percentage of German pronouns incorrectly translated in Europarl is the percentage of the total incorrect translations of *sie*, *er*, and *es* to the total of *sie*, *er*, and *es* in the test set.

Table 3.6: Overview of the annotations per error category in each test set for both English→Spanish and German→French. The error categories corresponding to content words (i.e. ambiguity, fluency, and misalignment) are subdivided into the part-of-speech common noun (N), adjective (Adj), and verb (V). The table shows the percentage (%) of the total of the current incorrectly translated group (i.e nouns, adjectives, verbs, connectives, or pronouns) (T) to its total in the test set.

		Europarl				WIT^3				News Commentary				Open Subtitles				Text+Berg	
		en-es		de-fr		en-es		de-fr		en-es		de-fr		en-es		de-fr		de-fr	
		T	%	*T*	%	*T*	%	*T*	%	*T*	%	*T*	%	*T*	%	*T*	%	*T*	%
Ambiguity	Noun	12	0.75	6	0.33	25	1.93	6	0.37	21	1.40	19	1.26	7	2.34	7	2.49	22	1.91
	Adjective	3	0.52	2	0.47	2	0.54	5	1.20	3	0.60	5	1.51	2	2.30	2	4.26	7	2.15
	Verb	15	1.09	7	0.39	9	0.93	15	0.90	9	0.83	7	0.69	4	1.67	2	0.68	4	0.86
	Total	30	0.82	15	0.44	36	1.37	26	0.70	33	1.07	31	1.09	13	2.08	11	1.77	33	1.70
Fluency	Noun	2	0.12	7	0.39	5	0.39	3	0.19	8	0.53	8	0.53	0	0.00	3	1.07	7	0.61
	Adjective	0	0.00	1	0.24	0	0.00	3	0.72	0	0.00	1	0.30	0	0.00	0	0.00	1	0.31
	Verb	2	0.14	0	0.00	1	0.10	5	0.30	3	0.28	0	0.00	0	0.00	1	0.34	1	0.21
	Total	4	0.11	8	0.24	6	0.23	11	0.30	11	0.36	9	0.32	0	0.00	4	0.64	9	0.46
Misalignment	Noun	6	0.37	9	0.50	4	0.31	14	0.87	7	0.47	13	0.86	1	0.33	5	1.78	19	1.65
	Adjective	1	0.17	1	0.24	1	0.27	4	0.96	3	0.60	4	1.20	4	4.60	0	0.00	5	1.54
	Verb	0	0.00	4	0.24	4	0.42	9	0.54	3	0.28	10	0.99	1	0.42	7	2.38	4	0.86
	Total	7	0.19	14	0.41	9	0.34	27	0.73	13	0.42	27	0.95	6	0.96	12	1.93	28	1.44
Connectives		3	1.39	0	0.00	1	0.68	4	2.80	1	0.98	2	2.67	0	0.00	1	6.25	0	0.00
Pronouns		8	10.39	12	13.64	2	2.74	53	26.5	2	4.00	23	35.94	0	0.00	9	16.67	3	16.67

TABLE 3.7: Overview of discourse (D) and local (L) annotations for **English→Spanish**. The total column *T* of each part-of-speech refers to the total of annotated errors of that part-of-speech in its error category (i.e. ambiguity or fluency) from table 3.6.

	Ambiguity									Fluency						**Total**		
	Noun			Adj.			Verb			Noun			Verb					
	L	*D*	*T*	*L*	*D*	*T*	*L*	*D*	*T*	*L*	*D*	*T*	*L*	*D*	*T*	*L*	*D*	*T*
EUROPARL	7	5	12	2	0	3	10	3	15	1	1	2	1	0	2	21	9	34
WIT^3	6	16	25	0	1	2	4	2	9	0	1	5	0	0	1	10	20	42
NEWS COMMENTARY	11	15	21	3	0	3	6	3	9	4	6	8	0	2	3	24	26	44
OPEN SUBTITLES	0	0	7	0	0	2	1	0	4	–	–	–	–	–	–	1	0	12
Total	24	43	65	5	1	10	12	8	37	5	8	15	1	2	6	**56**	**55**	**132**

TABLE 3.8: Overview of discourse (D) and local (L) annotations for **German→French**. The total column *T* of each part-of-speech refers to the total of annotated errors of that part-of-speech in its error category (i.e. ambiguity or fluency) from table 3.6.

	Ambiguity									Fluency									**Total**		
	Noun			Adj.			Verb			Noun			Adj.			Verb					
	L	*D*	*T*	*L*	*D*	*T*	*L*	*D*	*T*	*L*	*D*	*T*	*L*	*D*	*T*	*L*	*D*	*T*	*L*	*D*	*T*
EUROPARL	2	4	6	2	0	2	2	2	7	0	6	7	1	0	1	–	–	–	7	12	23
WIT^3	2	3	6	3	0	5	5	4	15	1	0	3	1	0	3	2	1	5	14	8	37
NEWS COMMENTARY	4	10	19	4	0	5	2	4	7	0	5	8	1	0	1	–	–	–	11	19	40
OPEN SUBTITLES	2	0	7	2	0	2	1	1	2	1	1	3	–	–	–	0	0	1	6	2	15
TEXT+BERG	10	4	22	1	1	7	3	0	4	5	0	7	1	0	1	0	0	1	20	5	42
Total	20	21	60	12	1	21	13	11	35	7	12	28	4	0	6	2	1	7	**58**	**46**	**157**

We observe that discourse errors, which represent less than 1% in most of the error categories, are not as frequent as other translation errors reported by Vilar et al. (2006) for English→Spanish. Specifically, Vilar et al. (2006) found that errors caused by bad tense amount to 15.1% of the total, followed by local reordering of words (11.6%), and missing content words (7.9%). These errors are not considered in our annotation task, since they are not discourse-related.

In our error annotation, pronouns result as the most problematic category, specially for German→French. Indeed, pronouns pose a challenge for MT and have received special attention in recent pronoun-translation competitions (Hardmeier et al., 2015, Guillou et al., 2016, Loáiciga et al., 2017).

We consider the German pronouns *sie*, *er*, and *es*, but we only find incorrect translations of *sie*. Figure 3.3 shows the relative frequency distribution of the translations of the German *sie* among the French pronouns *il*, *elle*, *ils*, *elles*, and *vous*, when it is incorrectly translated. We notice that the most frequent incorrect translations of *sie* produced by our German→French system are *vous* and *ils*. In example 3.6 we see that *sie* is incorrectly translated into *ils* ("they") instead of *vous* ("you" formal). Note that since we lowercased all data, we cannot keep the distinction between *Sie* ("you" formal) and *sie* ("they" or "she").

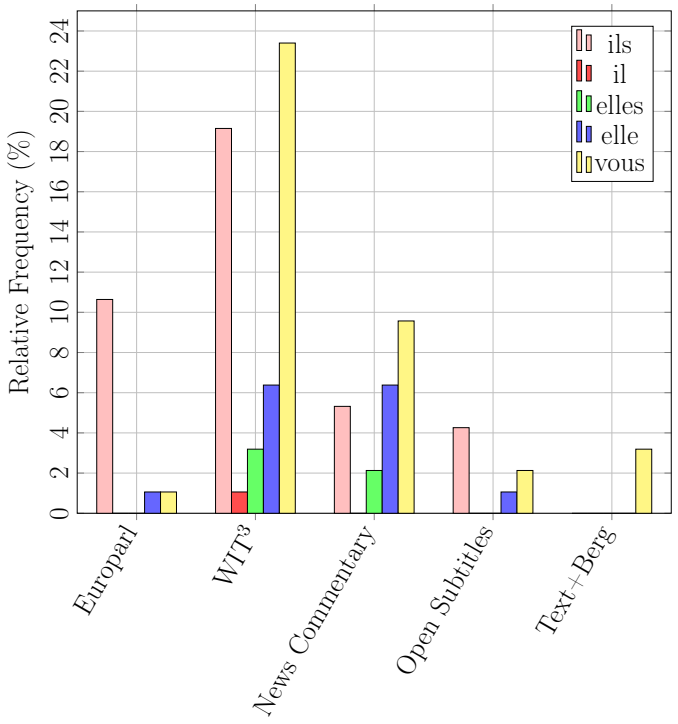


FIGURE 3.3: Percentage of the relative frequency of the incorrect translations of the German pronoun *sie* into the French pronouns *il*, *elle*, *ils*, *elles*, *vous* in each test set. The table shows that *sie* is most frequently mistranslated into *vous* and *ils*.

(3.6) **Source:** Reduzieren Sie ihre CO2-Emissionen durch jede Wahl, die *Sie* treffen können.
SMT: Réduire leurs émissions de CO2 par le choix, *ils* peuvent se rencontrer.

We find the highest percentage of wrong pronouns for English→Spanish when translating the Europarl test set (10.39%). Here, most of the errors are due to the mistranslation of the pronoun into its informal form. See example 3.7, where *you* is translated into the informal *ti* instead of the formal *usted*.

(3.7) **Source:** Mr Cox, Mr Hänsch, would this be acceptable to *you*?
SMT: Señor Cox, Sr. Hänsch, ¿esto sea aceptable para *ti*?

As for content words, ambiguity errors occur more often than fluency and misalignment errors in both German→French and English→Spanish (see also figure 3.4). Additionally, common nouns are translated into the wrong sense more often than adjectives and verbs for English→Spanish among all test sets (see figure 3.5). In example 3.8, the English *bills* in the sense of *legislative proposal* is incorrectly translated into the Spanish *facturas* ("invoice") in a document from the News Commentary about politics. Since another occurrence of *bills* in the same document is correctly translated into *proyectos*, we could solve this issue by encouraging the repeated use of the correct translation throughout the document.

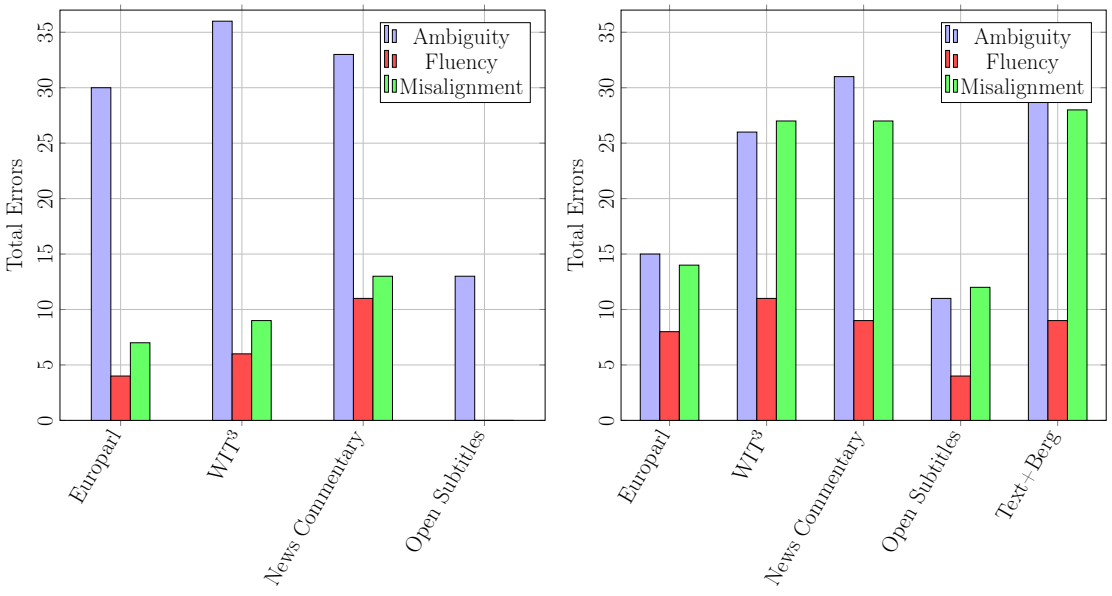


FIGURE 3.4: Total of fluency, ambiguity, and misalignment annotations for English→Spanish (left) and German→French (right). Ambiguity errors concern the majority of content words errors in both language directions.

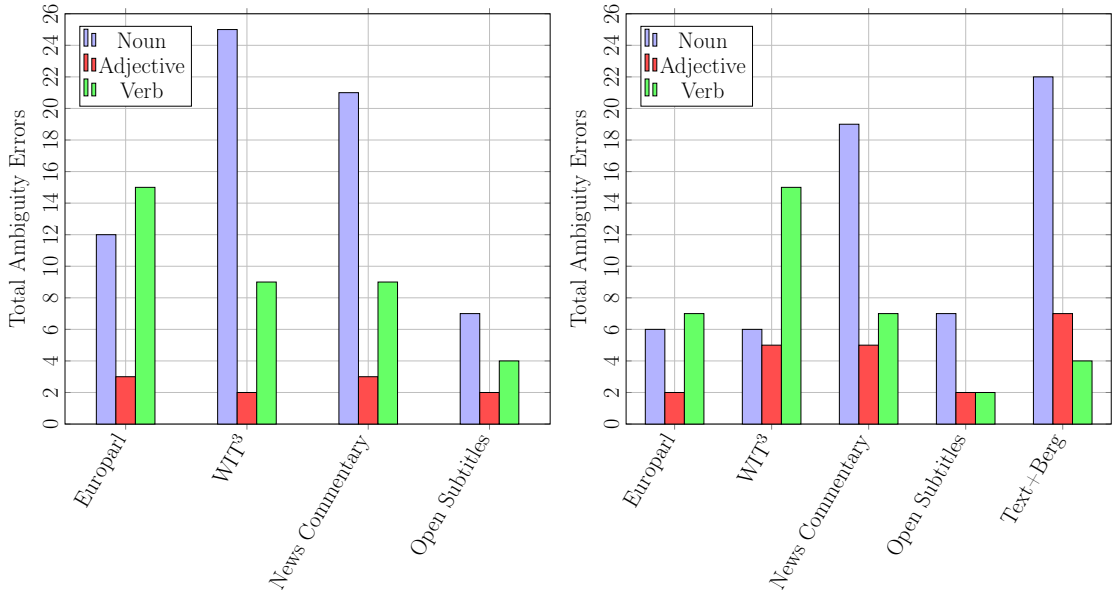


FIGURE 3.5: Total of ambiguity errors among common nouns, adjectives, and verbs for English→Spanish (left) and German→French (right).

(3.8) **Source:** As a result, 180 *bills* restricting the exercise of the right to vote in 41 states were introduced in 2011 alone.
SMT: Como resultado, 180 *facturas* restringir el ejercicio del derecho de voto en 41 países introducido en 2011.

Source: Several *bills* were blocked by vetoes of democratic governors.
SMT: Varios *proyectos* vetos de los bloquearon los gobernadores democrática.

Additionally, in example 3.9 from Europarl, we see that the English verb *appeal* in its present continuous form (*is appealing*) is incorrectly translated into the Spanish adjective *atractiva* ("attractive"). This issue can be solved at document level by keeping track of the words that define the topic of the document, such as *courts*, *acquitted*, *constitutional*, *right* and *prosecutor*, or at sentence level by training an SMT model that is able to distinguish between *appealing* as verb and as adjective. To do so, we need to train a model with part-of-speech tags and represent the test data in the same way. That is, each word is represented by its surface form and its part-of-speech as an additional *factor*. For example, the sentence *the public prosecutor is appealing* would be represented as *the|dt public|jj prosecutor|nn is|vbz appealing|vbg*.

(3.9) **Source:** All of us here are pleased that the courts have acquitted him and made it clear that in Russia, too, access to environmental information is a constitutional right.
Now, however, he is to go before the courts once more because the public

prosecutor is *appealing*.
SMT: Ahora, sin embargo, tiene que ir ante los tribunales, una vez más, porque el ministerio fiscal está *atractiva*.

In contrast to the English→Spanish translation, we find that the most ambiguity errors from German into French are produced by adjectives, followed by common nouns. In example 3.10 from Text+Berg, we see that *der Blick* ("the view") is translated into the wrong sense *le regard* ("the look").

(3.10) **Source:** Der *Blick* auf die Südwand des Bristen und ins Tal hinab ist atemberaubend.
SMT: Le *regard* sur la face du Bristen, et dans la vallée est effarant.

Fluency errors are considerably lower than ambiguity and misalignment errors, and we observe in both language pairs that they mostly affect common nouns. Since adjectives often go together with the noun that they describe, which provides local context, it is easier for the phrase-based SMT system to make good translation choices. As a result, we rarely find fluency errors from adjectives.

In example 3.11 from News Commentary, we see that the English *home* in *care home* is translated into the Spanish *hogar*, which would be rather used in the context of household. Indeed, it is the right translation for the first occurrence of *home* in the sentence. However, the preferred translation of *care home* is *centro de asistencia*, which refers to a residence for elderly or disabled people.

(3.11) **Source:** It is said that 77% of Canadians simply have no access to palliative care, which is care designed to ease the pain when a patient has reached the terminal stage of life, be it at home, in hospital or in a care *home*.
SMT: Se dice que 77% de los canadienses simplemente no tienen acceso a la atención médica paliativas, que está diseñada para aliviar el dolor, cuando un paciente terminal ha alcanzado el nivel de vida, ya sea en el hogar, en el hospital o en un *hogar*.

We also find verbs annotated as fluency errors. In example 3.12, the verb *pretend* would be better translated into Spanish *pretender* or *fingir* rather than *simular*.

(3.12) **Source:** We can no longer *pretend* not to understand this part of their suffering.
SMT: Ya no podemos *simular* no a comprender esta parte de su sufrimiento.

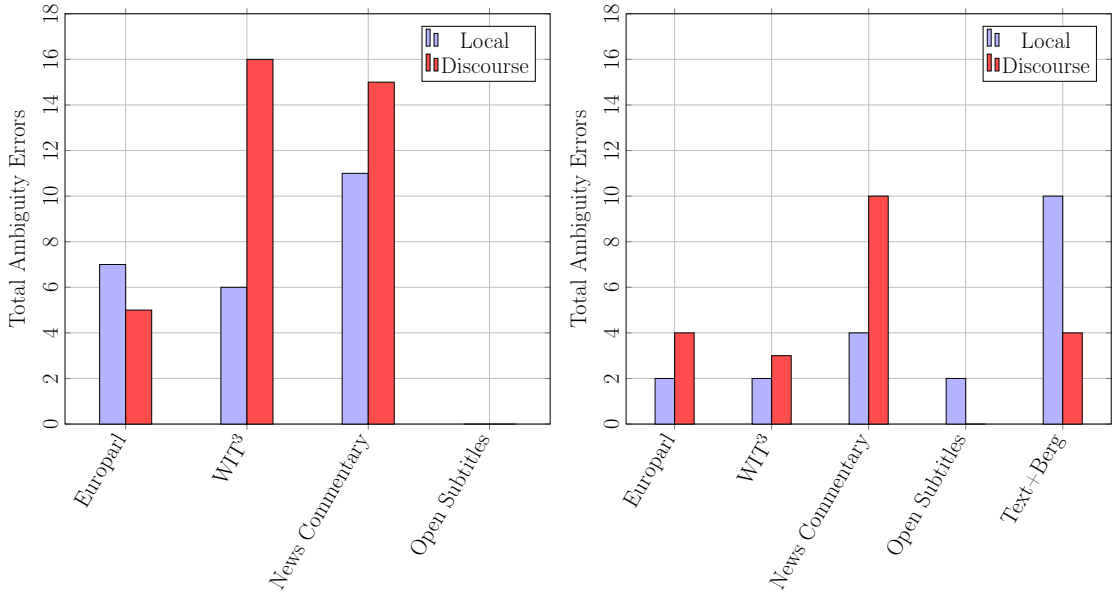


FIGURE 3.6: Total of ambiguous common nouns that can be solved using local or discourse context in each test set English→Spanish (left) and German→French (right). In both language directions, nouns benefit more from discourse context for disambiguation.

The last error category for content words concerns the misalignment errors. We find that these errors are more evenly distributed among common nouns, adjectives, and verbs than the other two categories aforementioned. See example 3.2 obtained from the WIT^3 test set.

For connectives we only find incorrect translations of the connectives *als*, *wenn*, and *wie* for German→French, and all the connective error annotations found in the WIT^3 concern to the connective *als*. We see in example 3.13 from WIT^3 the connective *als* mistranslated into the French *en tant que* ("as, like") instead of *quand* ("when"). In contrast, the majority of the connective errors in English→Spanish are in the Europarl test set. Example 3.14 shows the incorrect translation of the connective *since* (causal) into its temporal sense *desde*.

(3.13) **Source:** also, was habe ich *als* kleiner Junge getan?
SMT: donc, qu'est-ce que j'ai fait en tant que petit garçon?

(3.14) **Source:** My group believes that *since* a parliament is meant to listen, debate and reflect, there can be no justification whatsoever for this delay
SMT: A juicio de mi grupo, que *desde* el parlamento está destinado a escuchar, debate y reflexionar, no existe ninguna justificación alguna a este retraso

As described in the annotation guidelines (see section 3.1), we not only annotate the translation errors of content words, but also whether we could obtain a better lexical

choice by taking into account local context, discourse context, or both. As for discourse context, we distinguish three different categories: consistency, topic, or compounds.

We consider the compounds category specially for the mistranslation of common nouns from German into French, as German is a language rich in compounds. However, there is no noun labeled under this category in our test sets. The reason is that in those few cases where a noun is incorrectly translated and refers back to a compound, the compound is out-of-vocabulary and not translated. Therefore, the noun cannot benefit from the compound translation. See example 3.15 from WIT^3, where the German *Spots* refers to the compound *Fernsehspots* ("television advertisements"), but the compound is not translated by the German→French system.

(3.15) **Source:** Heute braucht man viele kurze, brandaktuelle, 30 oder 28 Sekunden lange *Fernsehspots*.
Wir müssen eine Menge solcher *Spots* kaufen.

SMT: Aujourd'hui, on a besoin d'une courte, de nombreux brandaktuelle ou 28, 30 secondes *Fernsehspots* depuis longtemps.
Nous devons acheter beaucoup de ces *points*.

Table 3.7 and table 3.8 summarise the local and discourse annotations in each test set for English→Spanish and German→French, respectively. We observe that in both language directions, nouns benefit more from discourse than local knowledge (see figure 3.6) and the reverse for adjectives. Indeed, adjectives usually go together with the noun they describe, which provides local context for disambiguation. Verbs benefit from local context and slightly less from discourse.

The total of discourse and local annotation in all test sets for each language direction is quite even, and local context is only slightly higher for German→French. This shows us that discourse knowledge is as important as local for the SMT system to improve the translation of the annotated errors.

3.4 Summary

In this chapter, we performed and analysed discourse errors in the translation of different types of text, such as news articles, movie subtitles, and transcribed talks. Specifically, we annotated errors in the translation of pronouns, connectives and content words (i.e. common nouns, adjectives, and verbs). We focused on the German→French and

English→Spanish and built the corresponding out-of-domain phrase-based SMT systems that we used to translate the test sets.

Content words that are incorrectly translated are labeled as one of the following categories that refers to the type of error: ambiguity, fluency, and misalignment. Additionally, we also annotated whether local or discourse knowledge could improve their translation. The latter is also subdivided into three different categories that refer to how the system could benefit from discourse: consistency, topic, and compounds.

The results of the analysis showed that ambiguity errors occur more often than fluency or misalignment errors. Furthermore, the ambiguity errors found for English→Spanish involved mostly nouns, whereas for German→French, we found mostly adjectives. In contrast, fluency errors are the least problematic and they concerned mainly nouns in both language directions.

We evaluated the results regarding whether discourse or local context could improve the translation of content words and observed that while nouns benefit more from discourse than local context, the opposite is true for adjectives. In general, both discourse and local can equally help to improve the translation of content words, which shows the importance of considering discourse knowledge in translation.

As described in chapter 2, the translation of pronouns has been extensively addressed in Hardmeier (2014)'s work and in a series of pronoun-prediction competitions. Similarly, Meyer (2014) performed a deep analysis of the translation of discourse connectives in his Ph.D. thesis. Therefore, we do not address pronouns and connectives in this thesis and only consider them in this chapter to get a better comparison of the main discourse errors. In the remaining of this thesis, we focus on content words and propose methods that benefit from the discourse context to improve their lexical choice.

Chapter 4

Consistency, or No Consistency

In this chapter, we tackle consistency in MT and analyse the controversial question of whether consistency is desirable in automatic translations. That is, if a word occurs more than once in the source document, do we want to use the same translation for all of the occurrences? Or should we encourage systems to introduce lexical variability instead?

We address consistency in the translation of two specific cases: references to a compound and pairs of repeated nouns. Compounds are words consisting of multiple morphemes, which can be referenced by its nominal head. For example, the German noun *Amt* ("office") and the compound *Bundesamt* ("federal office") can co-refer in a document.

In this chapter, we present two different approaches: `Post-editing`, and `Decode`. While the latter plugs the translation into the decoder, the former edits the translation output automatically. We assess the performance of the methods on a German→French system in the mountaineering domain, and a Chinese→English system built from transcriptions of the TED talks (section 4.2).

We then extend this consistency issue to any pair of occurrences of the same noun in a document, and build classifiers that predict whether they should be translated consistently, and if so, which translation should be used. To obtain a better comparison of the results between language pairs, we keep German and Chinese as source languages, and choose English as target in both cases. Furthermore, we use transcriptions of the TED talks to train both systems. Here, we use automatic post-editing and re-ranking to integrate the method with the SMT system, both in isolation and combined (section 4.3).

4.1 Introduction to Consistency

Consistency has been used in the literature to improve lexical choice in SMT. The main idea arises from the one-sense-per-discourse hypothesis, which states that multiple senses of the same word are not likely to occur in the same document (Gale et al., 1992). That is, for example, if we find multiple occurrences of the polysemic German word *Absatz* in a document, all of them are only in one of its senses: *heels*, *paragraph*, or *sales*.

This hypothesis extended later to MT with the one-translation-per-discourse hypothesis, claiming that consistency in translation is desirable (Carpuat, 2009). Carpuat and Simard (2012) show that human translators and SMT systems translate remarkably consistently, and that inconsistencies signal translation errors more often than consistencies do. However, they also reveal that consistency is not a good indicator of translation quality. Translation systems trained on large text collections deal with more translation choices, and therefore, they translate more inconsistently.

Repetition as a consequence of strict consistency enforcement is also discussed, since it is difficult to determine whether repetition is desirable or not (Carpuat and Simard, 2012). On the one hand, Carpuat (2009) shows that human translators tend to use repetition across the document. On the other hand, it may negatively affect fluency (Guillou, 2013).

Instead of tackling translation consistency in general, in this chapter we focus on two special scenarios: references to noun compounds and pairs of repeated nouns. The idea is to enforce consistency in cases where we assume that a repeated translation is expected, and to evaluate how this affects the translation quality. This way we intend to avoid overusing consistency, and negatively affecting the fluency of the translation.

4.2 Nominal References to Noun Compounds and Consistency

The nominal head of a noun compound can be used in subsequent sentences to co-refer the same entity. For example, the German compound *Nordwand* ("North face") and its nominal head *Wand* can co-refer in a document. Intuitively, the nominal head should have the same translation in the compound and in its references. However, since SMT systems translate at sentence level, this nominal head may be translated inconsistently across the document. This inconsistent translation leads to errors when the nominal head has several translations in the target language. Consider the following example:

(4.1) . . . on the unclimbed East face of the Central Tower . . .
. . . we were swept from the *face* by a five-day storm . . .

The English word *face* is most frequently translated into German as *Gesicht* ("front head"), but in the example 4.1, *face* refers to a side of the Central Tower ("East face") and must be translated into German as *Seite*. Accordingly, the SMT system needs to consider the context in order to determine the correct translation variant.

In this section, we address consistency on references to a compound. We tackle a specific case in which the compound is referenced using its nominal head, proposing a method to consistently translate those references using the corresponding translation from the compound. The idea is that compounds have less translation variants than single morpheme words, and therefore there is less ambiguity in translation. For example, while the English word *face* can be translated into German as *Gesicht* or *Wand*, there is no term such as *Ostgesicht* in German. Only German words in the sense of side wall such as *Ostseite* can be used to translate *East face*.

4.2.1 Automatic Detection of References to Compounds

A compound can be referenced by its nominal head. For example, the compound *Nordwand* ("north face"), formed by *Nord* (X) and *Wand* (Y) can be referenced by *Wand* (Y). Compounds can also consist of more than two morphemes, and their nominal head can also be composed by multiple lexemes. For instance, the head of the compound *Eiger-Nordwand* ("Eiger north face") can be either *Wand* or the compound *Nordwand*. The main aim of our method is to detect references to noun compounds, where the reference is the nominal head of the compound (Y), and to enforce Y to have the same translation in both the compound (XY), and the reference (Y).

To consistently translate Y, we cache (i.e. store) its translation from the compound and enforce it when a reference to it is detected. In general, caching is sensitive to error propagation as Tiedemann (2010) points out. That is, if we cache an incorrect translation, it then gets propagated throughout the document. However, the scope of our approach is narrower than Tiedemann (2010)'s approach as it does not consider any content word, but only compounds, which provide more context for a correct translation than single morpheme words, yielding less translation variants.

To identify compounds, we first analyse each noun with the German morphology system *Gertwol* (Koskeniemmi and Haapalainen, 1994), which marks the boundaries between independent morphemes. For instance, the analysis of *Ostwand* is *Ost#wand*. We then

obtain the translation of the compounds, and cache them. To automatically obtain their translation, we use the word alignments generated at the decoding step.

To identify Y as a reference to a previously seen compound, we apply the pattern *determiner + (adjective) + Y lemma*, where the *adjective* is optional, and the *determiner* is tagged as one of the following parts-of-speech: (1) PDAT or attributive demonstrative pronoun (e.g. *jener*), (2) PPOSAT or attributive possessive pronoun (e.g. *mein*, *deine*), and (3) ART or article (e.g. *der*, *die*, *das*). Note that we only consider definite articles as indefinite articles such as *ein* (English "a" or "an") are used to introduce new entities, and do not refer to a previous mention.[1]

Thus, *die prächtige Fahrt* ("the magnificent ride") and *diesen Grat* ("this ridge") are examples matching the pattern. We use the lemma of Y to also match examples where German cases (e.g. genitive and dative) change the form of Y (e.g. *Grates* is the genitive form of *Grat*). We then check that a compound XY and a reference Y are in a four-sentences window, since a larger window introduces too much noise. Exceptionally, we consider the whole document when the determiner is PDAT (e.g. *diese*). PDAT is a strong reference indicator, and we found examples having more than four sentences between the compound and its reference.

Different occurrences of Y can vary their translation depending on the compound to which they belong. For instance, the noun *Wand* ("wall") in *Nordwand* ("north face") and *Felswand* ("rockface") are often translated into French as *face* and *paroi*, respectively. If there are several compounds sharing the head, the method needs to decide to which of them Y refers to. Intuitively, Y refers to the closest one, but not necessarily. To better detect the antecedent of Y, the method should analyse the context, but for the sake of simplicity, we assume in these experiments that Y refers to the last matching compound translated.

4.2.2 Integration of the Consistency Method with the SMT System

In this section, we explore three different approaches to enforce the correct translation of Y, and evaluate how they perform at detecting the translation of Y from the translation of the compound. The first two approaches tackle the translation of Y at the decoding stage, plugging into the decoder the translation that must be used. We refer to the technique of plugging the cached translation into the decoder as `Decode`. The third approach

[1] In these experiments, we did not consider the part-of-speech APPRART (e.g. *im*, or *zur*), which is a contraction of a preposition with a definite article. After analysing the results we noticed that it could improve recall, and it should therefore be included in future experiments.

consists of automatically editing the output of the translation with the cached translation (`Post-editing`). In the rest of this section, we describe the different approaches in detail.

In the first approach, we let the decoder decide which is the best translation of Y from the compound translation. To do so, the method first caches the translation of a compound XY as a translation of Y. Next, when a reference Y is detected, the method plugs all the content words cached into the decoder without assigning them any probability. As a consequence, every plugged word is equally probable to be the correct translation. The decoder then chooses the best candidate based on translation and language model scores.

In the following command line we run the decoder, and tackle the translation of *Typ* ("type" or "guy"), which refers back to *Körpertyp* ("body type"):

```
$ echo 'Der ektomorphe Körpertyp neigt zur Schlankheit, deswegen muss \
  dieser <n translation="body||type">Typ</n> viel Krafttraining machen' \
  | moses -xml-input exclusive -f moses.ini
```

In this example, we use the flag *-xml-input* to tell the decoder Moses that we want to use a specific translation for a word or phrase. We then must use a XML markup scheme to specify a translation as in the example. The value of the flag *exclusive* tells the decoder that the specified translation must be used for the input phrase. Moses has other value flags available such as *inclusive*, and *constraint*. These allow the specified translation to compete with other translations in the phrase table. Since we do not have a probability for the suggested translation, the decoder often ignores it, and picks the most likely translation from the phrase table. We did not notice an advantage in using *inclusive* or *constraint* without providing a probability, and we therefore force the decoder to use the specified translation with *exclusive*.

The translation of the compound must always be aligned to more than one word on the target side. If it is only aligned to one word it might be due to misalignment or lexicalisation of the compound. A lexicalised compound cannot be referenced by its head, since its translation does not correspond to the translation of its components. For example, the German compounds *Zusammenarbeit* ("cooperation") and *Augenblick* ("moment") are lexicalised and thus, they cannot be referenced by *die Arbeit* ("the work") or *der Blick* ("the view"), respectively. In such cases, the method does not proceed with the approach to avoid caching and propagating a wrong translation.

Interestingly, this approach fails in the experiments. We observe that the decoder takes the translation of the constituent that appears more frequently in the language model independently on whether it is the head of the compound or not. For example, *Wand* as a reference of *Nordwand* ("north face") is enforced to be translated into *north*. In this

case, the score computed for the first constituent is higher, and it is then picked as the translation candidate.

In the second approach, for each content word of a compound translation, the method checks whether it appears as a translation candidate of Y in the phrase table. The method then caches only the translation candidate that has the highest direct phrase translation probability. Since the method only considers a translation if it is in the phrase table, it can also work when the compound is aligned to only one word in the translation.

We observe that by applying this method, some test examples where the compound was aligned to only one word in the target side are improved. In the first approach, we assume that compounds aligned to only one word may force the reference to use an incorrect translation, so they are detected as false positives and discarded. However, in this second approach, the translation is used when it appears as a translation candidate of Y in the phrase table, resulting in a better translation of the term according to the context. Therefore, the German word *Fahrt* is translated into *ascension*, which is also the translation of the referenced compound (*Bergfahrten*) as shown in example 4.2.

(4.2) **Source**: Unter den Neuen ***Bergfahrten*** [*ascension*] in den Schweizeralpen ist im IV. Band der Alpen 1928 eine erste Begehung des ganzen Südostgrates von der Gamsenlücke ...
über die prächtige ***Fahrt*** geblieben.

English Human Translation: Among the new *Alpine hikes* in the Swiss Alps, the first inspection of the entire southeast ridge of the Gemsenlück is mentioned in the IV. volume of the Alps 1928 ...
remained about the magnificent *journey*.

Baseline, Baseline$_{split}$: par cette magnifique ***course***.
`Cpd`, `Cpd`$_{split}$: par cette magnifique ***ascension***.

At decoding stage, the Moses decoder must find the highest scoring translation of a given sentence. This score is computed taking into account individual probabilities from each model (e.g. translation and language model). Thus, to enforce a specific translation during decoding is usually preferable than modifying the translation afterwards, since it then considers the model scores. However, the approach of plugging the translation into the decoder is not optimal as the enforced translation is introduced without probability scores (Carpuat, 2009). We therefore also try an automatic post-editing approach, which

modifies the translation output of the references with the cached translation through word alignment.

We finally conducted the experiments with the last two approaches. The method then checks the phrase table to obtain the best translation candidate, and either plugs it into the decoder (section 4.2.3) or automatically post-edits the translation output with the translation cached (section 4.2.4). Both experiments are carried out using a German→French SMT system in the mountaineering domain. The results are also compared with the output of a Chinese→English SMT system, which uses transcribed texts for training and testing.

4.2.3 Consistent Translation of References to Compounds from German into French

The translation of compounds is the first step to proceed with our method. However, compounds are often out-of-vocabulary (i.e. they do not appear in the training corpus) and the system cannot translate them. These compounds are usually composed of frequent words in the training corpus, so we can obtain the translation of an unseen compound by splitting it into its known parts and translating them (Koehn and Knight, 2003). We want to assess the performance of our method in both approaches (i.e. splitting compounds and not splitting them), so we build two phrase-based SMT systems `Cpd` and $\texttt{Cpd}_{split}$, where the latter performs compound splitting.

The data comes from the Text+Berg corpus (see section 1.3.5). We train the language model on a total of 624,160 sentences (13 million target tokens) and the translation systems on 219,187 sentences (roughly 4.1/4.7 million words) in German and French, respectively. The SMT systems are tuned on a development set, also from Text+Berg, consisting of 1,424 sentence pairs and approximately 31,000 tokens for each language. The test set is a collection of 318 examples, that is, groups of sentences containing a compound noun and its references, randomly sampled from Text+Berg data.

We expect to enforce a consistent translation in a higher number of cases with the $\texttt{Cpd}_{split}$ system. Furthermore, the splitting method allows us to have a one-to-one alignment between the compound constituents and their translation. Thus, we can identify the translation of the head of the compound and cache it directly.

4.2.3.1 Evaluation of the Automatic Detection of References to German Compounds

To evaluate how often a German compound is referenced by its nominal head, we automatically detect them in a German corpus consisting of roughly 1.1 million sentences from Text+Berg as described in section 4.2.1. As a result, we obtain 24,317 instances.

Two annotators conduct then a manual analysis of a random sample containing 318 compound-reference pairs automatically detected. The task consists on annotating whether the detected reference and the compound co-refer, and if so, whether the translation of the reference is correct or not. In those cases where the systems produce different translations, they must annotate the quality of the translation for each of them. The annotators do not know the difference between systems, and which of them produces each translation. The agreement between them at the task of deciding "is/is not a co-reference" and "is correct/wrong reference translation" is 73.4% and 86.8%, respectively.

Example 4.3 illustrates the format of the annotation task that the annotators receive for each reference to a compound detected. The numbers 1 and 2 stand for the baseline systems, where 2 performs compound splitting. Similarly, 3 and 4 are the `Cpd`, and $\texttt{Cpd}_{split}$ systems, respectively. The annotators are not informed of their meaning.

(4.3) **Compound**: Gipfelkrater

Context: Nebst landschaftlicher Vielfalt ist die Übernachtung im Zelt im *Gipfelkrater* auf 5800m Höhe das Highlight.
Context Automatic Translation: nebst est la diversité du Übernachtung dans la tente le à 5800m, le highlight.[2]

Source: Damit verbunden ist eine Erkundung des *Kraters* mit seinen imposanten Gletschern
(Q) Is it a compound coreference?: [+/-]

Automatic Translation (1 2): ainsi est une reconnaissance du cratére, avec ses glaciers
(Q) Is the coreference translation ok?: [+/-]

Automatic Translation (3): ce n'est pas une reconnaissance du cratére,

[2]The system did not translate the German phrase *landschaftlicher Vielfalt*.

avec ses imposantes des glaciers
(Q) Is the coreference translation ok?: [+/-]

Automatic Translation (4): ce n'est pas une reconnaissance du cratére avec ses imposantes des glaciers
(Q) Is the coreference translation ok?: [+/-]

The manual analysis reveals that 107 of these pairs are false positives. These false positives are due to the lexicalisation of the compound or a number disagreement between the compound and its reference. Example 4.4 shows an incorrect reference to *Zusammenarbeit* ("cooperation") as it cannot be referenced by its nominal head *Arbeit* ("work").

(4.4) **Source**: Du erlebst hautnah, was ein sonniger Tag an Hektik bringt und wie wichtig eine gute ***Zusammenarbeit*** im Hüttenteam ist.
Anders als bei Work&Climb steht hier die ***Arbeit*** im Vordergrund, denn du bist eine wertvolle Arbeitshilfe für den Hüttenwart.

English Human Translation: You will experience first-hand what a sunny day brings to the hustle and bustle and the importance of a good ***cooperation*** in the Hüttenteam.
Unlike Work&Climb, work is the focus, because you are a valuable ***work*** aid for the hut keeper.

We could avoid false positives due to number disagreement between the compound and its references by ignoring references that do not have the same grammatical form than the head of the compound. However, this would not allow the method to detect different forms of Y due to the German grammatical cases. For this reason, the method matches the lemma instead, allowing us to detect, for example, *Firngrat* ("firn ridge") and *Grates* ("ridge") as co-referent, where the latter is in its genitive case form (see example 4.5).

(4.5) **Source**: Die Punta Isabella entsendet nach Süden einen Feis- und ***Firngrat***, der mit einer Steilwand im Trioletgletscher fußt.[3]
In der Westilanke dieses ***Grates***, knapp oberhalb des Abbruches über eine Schneezunge hinauf und nach rechts in die Felsen.

English Human Translation: The Punta Isabella sends a rock and ***Firn***

[3] *Feis* is an OCR error that corresponds to *Fels* ("rock")

> ***Ridge*** to the south with a steep wall in the Triolet glacier.
> In the west bank of this ***ridge***, just above the demolition over a snow tongue and to the right in the rocks.

In other less frequent cases, the detected coreference has nothing to do with the compound. For instance, in the example 4.6, the pattern correctly matches the reference *Gipfel* ("summit"), but the method fails at detecting *Schneegipfel* ("snowy summit") as the compound referenced. Indeed, *Gipfel* ("summit") refers to the mountain *Königsspitze*.

(4.6) **Source**: Er sah von ihr wirklich auf den obern Trafoierferner links hinunter und erblickte über mehrere ***Schneegipfel*** hinweg sein Ziel, die im Hintergrunde sich erhebende Königsspitze. Auf deren ***Gipfel*** grub er sich dann halbliegend in den zusammengewehten Schnee ein.
English Human Translation: He indeed looked down to the left on the upper Trafoierferner and saw his goal over several *snow peaks*, the *Königsspitze* rising up in the background. On its *summit*, he buried himself half-way into the snow dunes.

The manual analysis also focuses on the correct detections, distinguishing the following most common patterns:

- The reference is preceded by a definite article and an adjective or by the demonstrative adjectives *dieser* ("this") and *jener* ("that") in all their grammatical forms.
- The compound is in genitive case and its reference in nominative or dative case. For example, *das Tal* ("the valley") refers to *Haupttals* ("main valley") in *Sohle des Haupttals* ("bottom of the main valley").

4.2.3.2 Manual Analysis of the Decode Approach

In this section, we present the results on both correctness and consistency for both systems `Cpd` and $\texttt{Cpd}_{split}$. The results are compared against a baseline, which does not influence the translation in any way. The experiments are performed on the 211 compound-coreference pairs correctly detected in the original test set. Example 4.7 shows the improvement of our method. Here, we observe that the correct translation of the German *Amt* is enforced by `Cpd`, and $\texttt{Cpd}_{split}$, whereas both baselines incorrectly translate it into *post*.

(4.7) **Source**: Die Originalauswertung wurde in den Zwischenmassstab 1:20000 reduziert, worauf das ***Bundesamt*** [*office fédéral*] für Landestopographie in Aktion trat.
Nur dieses ***Amt*** war in der Lage, ...

English Human Translation: The original evaluation was reduced in the intermediate scale 1:20000, after which the *Federal Office* of Topography came into action.
Only this *office* was able to ...

Baseline, Baseline$_{split}$: que ce ***poste*** était dans la situation, ...
Cpd, Cpd$_{split}$: que de cet ***office*** était en mesure ...

The precision of our method at correctly detecting a reference to a compound is 66.4% (i.e. 211 out of 318 coreferences). In example 4.8 our method incorrectly detected *Westseite* ("west side") as reference of *Nordwestseite* ("north west side"). In the first sentence, the compound refers to a side of the amphitheater, whereas in the second sentence, the incorrectly detected reference concerns a side of the Sentinel.

(4.8) **Source**: Der interessanteste Kletterberg auf der ***Nordwestseite*** des Amphitheaters ist der Sentinel.
Soweit sind uns die Fakten bekannt, als wir am 25. Februar ausrücken, um den Sentinel über seine ***Westseite*** zu besteigen.

English Human Translation: The most interesting climbing mountain on the ***north west side*** of the amphitheater is the Sentinel.
So far as the facts are known to us, on February 25, we are going out to climb the Sentinel over his ***western side***.

There are no incorrect translations enforced with our method in the cases where the reference is incorrectly detected. This is due to the fact that the method only enforces a reference translation when the translation candidate is in the phrase table. Since we only analyse the sentences detected by the method, recall is not computed. However, our detector's approach is broad-coverage-oriented. That is, it tends to detect more false positives while practically avoiding false negatives.

To analyse the coverage of the method, we consider not only the correctly detected references, but also the false positives. Specifically, 42.2% (i.e. 89 out of 211) of the positive examples, and 27.1% (i.e. 29 out of 107) of false positives enforce a reference translation when compound splitting is not performed. The remaining 57.8% of the positive

TABLE 4.1: Results on consistency and correctness for the baseline and `Cpd` systems.

	Consistent		Inconsistent	
	Correct	Incorrect	Correct	Incorrect
BASELINE	52	6	117	36
CPD	73	7	102	29

TABLE 4.2: Results on consistency and correctness for the baseline$_{split}$ and `Cpd`$_{split}$ systems, where both perform compound splitting.

	Consistent		Inconsistent	
	Correct	Incorrect	Correct	Incorrect
BASELINE$_{split}$	68	6	105	32
CPD$_{split}$	103	6	80	22

examples (i.e. where no enforcing is applied) are due to out-of-vocabulary compounds and misalignments. Splitting significantly increases the coverage of enforced translations from 42.2% to 56.4% (i.e. 119 out of 211). The incorrectly identified references have again a lower impact ratio (34.6%; 37 out of 107).

The baseline system correctly translates with 80.1% accuracy and 27.5% consistency (see table 4.3). The German noun *Wand* is the most common example of inconsistent but correct translation in our test set. The most likely translation for this noun in French is *paroi* in the Text+Berg corpora. However, when *Wand* is part of a compound, it is usually translated into the French *face*.

Table 4.1 shows the results of `Cpd` compared to the baseline. The system enforces a consistent translation in 89 of the cases improving the translation of six of them. Moreover, 15 test pairs stay correct, but become consistent. For instance, in example 4.9, the noun *Gebiet* ("area") is translated into *site* instead of *région* when a consistent translation is enforced, yet both translations are correct.

(4.9) **Source**: Dass dies gemacht wird, zeigt das Routenbuch "Clean-Begehungen", das im ***Klettergebiet*** [*site d'escalade*] liegt.
Wir diskutieren über die schönsten Routen im ***Gebiet***.

English Human Translation: That will be done, the route book "Clean-Begehungen" shows that it is in the *climbing area*.
We discuss about the nicest tours in the *area*.

Baseline, Baseline$_{split}$: nous discutons sur les plus belles voies dans la ***région***.
Cpd, Cpd$_{split}$: nous discutons sur les plus belles voies du ***site***.

We also observe in table 4.1 that only one reference to a compound becomes consistent while still being incorrect. The remaining 67 stay unmodified, that is, the baseline system already translates the reference consistenly. Overall, while the correctness of Cpd is slightly raised from 80.1% to 82.9%, the consistency improves from 27.5% to 37.9%.

The results of the method are improved when we perform compound splitting on the data. Specifically, only three cases become worse, and most of the cases that are not enforced with the Cpd system due to misalignments or out-of-vocabulary compounds, are now enforced and improved. For instance, in example 4.10 we observe that the Cpd system does not enforce any translation of the German noun *Heft*. That is because *Quartalsheft* is misaligned to only *trimestrel*, which does not appear as a translation candidate of *Heft*, and it is therefore not enforced. However, when applying the compound splitting technique, there is one-to-one correspondence between *Heft* and *numéro*. The coreference translation is then successfully enforced by the system Cpd$_{split}$.

(4.10) **Source**: Einen Teil ihrer bergsteigerischen und wissenschaftlichen Erfolge finden unsere Mitglieder in diesem ***Quartalsheft*** [*présent numéro trimestriel*] verzeichnet.
Das vorliegende ***Heft*** möge daher ...

English Human Translation: Our members find part of their mountaineering and scientific achievements in this quarterly *quarterly bulletin*. This *bulletin* may therefore ...

Baseline, Baseline$_{split}$**, Cpd**: le ***cahier*** möge donc ...
Cpd$_{split}$: le présent ***numéro*** möge donc ...

The system baseline$_{split}$, which does not enforce any reference translation, translates correctly 82.0% of the test pairs, and consistently 35.1% of them. When enforcing Cpd$_{split}$ achieves 86.7% correctness, and 52.2% consistency (see table 4.2).

Compound splitting increases the coverage of the method, and improves the translation output as is shown in table 4.2. Indeed, it applies enforcing to 109 cases improving 10 of them. Although there are six consistent and incorrect cases in both table 4.1 and table 4.2, some of them are different. Specifically, Cpd$_{split}$ improves two of them and makes consistent another two, although both stay incorrect. The correctness rises from

TABLE 4.3: Overall percentages of consistency and correctness results of `Cpd` and $\texttt{Cpd}_{split}$ systems, with and without applying our enforcing method.

	Correctness	Consistency
BASELINE	80.1%	27.5%
BASELINE_{split}	82.0%	35.1%
CPD	82.9%	37.9%
CPD_{split}	86.7%	52.1%

82.0% to 86.7% and consistency from 35.1% to 52.1%. Overall, the final effect is positive (see table 4.3). Correctness rises from 80.1% to 86.7%, improving 17 examples (i.e. one third of errors are fixed), and consistency from 27.5% to 52.1%.

4.2.4 Comparison of Approaches: Decode versus Post-editing

In this section, we compare the `Decode` approach, which plugs the reference translation from the translation of the compound into the decoder, with `Post-editing`, which automatically post-edits the translation of the references to compounds. As stated in section 4.2.2, the `Decode` approach is not optimal, since it introduces translation candidates without probabilities that compete with the ones from the phrase table. `Post-editing` changes the translation output, avoiding this issue. However, since post-editing is done after decoding, the models integrated in the decoder, such as translation and language model do not take part in the process.

To get a better insight of the comparison, we carry out the translation experiments from German to French and also from Chinese to English. Chinese is a non-segmented language (i.e. words in a sentence are not separated by blank spaces) and therefore, the texts need to be segmented. After applying word segmentation using the Stanford Word Segmenter,[4] it is possible to distinguish multi-character words as in the German compounds. For example, in the words 高跟鞋 ("high heels") and 蔬菜 ("vegetables") each character has an individual meaning. These multi-character words can also be referenced using the last character. For example, 高跟鞋 ("high heels") and 鞋 ("shoe") can co-refer in the same document, and then the latter should be translated into *heels*.

Our data to train the Chinese→English system comes from the WIT[3]. We also use the Text+Berg corpus to train the German→French system (see section 1.3.5 for a description of the corpora). Even though the latter system is trained on the same corpus as in

[4] http://nlp.stanford.edu/software/segmenter.shtml

TABLE 4.4: Size of the data used to train the SMT systems that translate from German to French (Text+Berg), and from Chinese to English (WIT^3).

		Sentences	Tokens
ZH	Training	188,758	19,880,790
	Tuning	2,457	260,770
	Testing	855	12,344
DE	Training	285,877	5,194,622
	Tuning	1,557	32,649
	Testing	505	12,499

section 4.2.3, the systems are not trained and tested exactly on the same data. The reason is that in section 4.2.3 the test sets were extracted from all data available including the monolingual part. In these experiments, the test sets are different, and obtained only from parallel data. This way, it is possible to apply automatic measures such as BLEU to measure the quality of the translation. Table 4.4 details the sizes of both systems.

The test set is a collection of 261 compound-reference pairs in both language pairs. While our method detected a total of 7,365 pairs among 192k sentences in German, only 261 were detected in the Chinese data. The reason is that the pattern to detect references Y to a compound in Chinese is more restricted than in German: only the corresponding demonstrative pronouns in Chinese *this* and *that* in less than a three-words distance can precede Y. The 261 pairs of the German test set are selected randomly.

The evaluation of the results is done automatically in two different ways. We first use BLEU to compute the overall score of the translation output of the two different approaches and compare it with a baseline that does not enforce any translation reference and an oracle translation. In the oracle translation, all translations of the references match with the human translation, which indicates the maximum BLEU score that it can reach for each language pair. To get a better insight, we then automatically compute through word alignment the total of references that match the human translation and present results accordingly.

The BLEU scores per language pair are listed in table 4.5, showing that in both languages `Post-editing` outperforms `Decode`. The results also reveal that while the two approaches

TABLE 4.5: BLEU scores of our methods

	ZH-EN	DE-FR
BASELINE	11.18	27.65
DECODE	11.23	27.26
POST-EDITING	11.27	27.48
ORACLE	11.30	27.80

TABLE 4.6: Comparison of each approach with the baseline for Chinese→English and German→French. The table shows the percentage of Y that match or differ from the human translation (ref). The numbers in **bold** are improvements over the baseline, and those in *italics* are degradations.

			DECODE		POST-EDITING	
			= ref	≠ ref	= ref	≠ ref
ZH-EN	BASELINE	= ref	59.3	*4.1*	42.3	*4.5*
		≠ ref	**13.8**	22.8	**20.3**	32.9
DE-FR	BASELINE	= ref	70.1	*10.3*	73.9	*5.0*
		≠ ref	**4.3**	15.2	**3.5**	17.5

(i.e. `Post-editing` and `Decode`) have a small positive effect on the translation from Chinese to English, they have a small negative effect on German→French. The Oracle scores show that even a perfect matching of all pairs with the reference would not have a big impact on BLEU scores.

Table 4.6 presents a more detailed analysis of the results. The table shows the percentage of improvements and degradations over the baseline. The `Post-editing` approach in Chinese→English has a net improvement of 15.8% as it improves 20.3% of the cases, and degrades 4.5%. This improvement is larger than with the `Decode` approach, which shows 13.8% improvement and 4.1% degradation over the baseline. In the German→French translation, both methods score fewer improvements than degradations. The baseline and the systems that apply the two approaches correctly translate more than 70% of the pairs, which indicates that the margin for improvement is much smaller for German→French than for Chinese→English.

In both language pairs, the `Post-editing` approach shows a larger improvement than `Decode` as the difference between improvements and degradations is larger. Note that the `Decode` approach only enforces a translation when it appears as a translation candidate of the reference in the phrase table. The coverage of `Decode` is therefore lower than in `Post-editing`, which explains the smaller improvement.

4.3 Consistent Translation of Repeated Nouns

In this section, we extend the topic of consistent translation of references to compounds, considering any pair of repeated nouns in a source text. Since repeated nouns in a text refer to the same entity (i.e. they co-refer), their translation must have the same sense. Note that a word might have several translations in a target language, some are synonyms, but others represent different senses. We therefore address this issue translating both consistently.

TABLE 4.7: WIT^3 data for building the SMT systems

WIT^3	MT training		MT tuning		Language modeling	
	Sentences	Words	Sentences	Words	Sentences	Words
DE-EN	193,152	3.6M	2,052	40K	217K	4.4M
ZH-EN	185,443	3.4M	2,457	54K	4.8M	800M

The following examples illustrate the presented issue. In example 4.11, the system incorrectly translates one of the occurrences of the German *Politik* into the English *politics*. Similarly, in example 4.12 the system translates the Chinese characters 证件 into *identity papers*, which is less idiomatic and frequent than *identity documents*.

(4.11) **Source**: nach Einführung dieser **Politik**... die **Politik** auf dem Gebiet der Informationstechnik...
SMT: after introduction of **policy**... the *politics* in the area of information technology...
Human Reference: once the **policy** is implemented... the information technology **policy**...

(4.12) **Source**: 欺诈性旅行或身份**证件**系指有下列情形之一的任何旅行或身份**证件**
SMT: 欺诈性 travel or identity *papers*. 系指 have under one condition; any travel, or identity **document**
Human Reference: Fraudulent travel or identity **document**; shall mean any travel or identity **document**

To tackle this issue, we train a set of classifiers on syntactic and semantic features that predict how to consistently translate a pair of nouns. Specifically, they predict whether the pair of nouns should be consistently translated, and if so, which of the translations from the two nouns should replace the other one. The experiments are performed on the Chinese→English and German→English translation. Note that in this section, the language pairs share the target language. This way, it is possible to get a better comparison between Chinese and German in this task.

The data for both Chinese-English and German-English comes from the WIT^3 corpus, and the UN Corpora, a collection of documents from the United Nations.[5] The SMT baseline systems (i.e. one per language pair) are built on WIT^3 data using Moses with

[5]http://www.uncorpora.org

TABLE 4.8: UN data to train and test the classifiers.

UN Data	Classifier training			Classifier testing		
	Sentences	Words	Nouns	Sentences	Words	Noun
DE-EN	150K	4.5M	11,289	7,771	225K	695
ZH-EN	10K	368K	3,301	3,000	121K	647

the defaults settings (table 4.7). The UN Corpora are then used for testing and training the classifiers (table 4.8).

4.3.1 Detection of Noun Pairs

To detect the noun pairs that are used to train and test our method, we proceed as follows. We extract all pairs of nouns occurrences (i.e. $N_1 \dots N_2$, where N_1, and N_2 are different occurrences of the same noun), whose SMT translations are different (i.e. $T_1 \neq T_2$). Our method cannot improve the translation of pairs of nouns that are already consistently translated.

Our classifiers are build on supervised training data. The classifiers need to predict whether the noun pair should be consistently translated (class: *None*), and if that is the case, which of the two occurrences is correctly translated and should substitute the other. Accordingly, a training instance gets labelled as class 1, when the first occurrence is already correctly translated, and 2 otherwise.

We therefore label the training data with the correct prediction. To do so, we rely on the reference translation, which we obtain through word alignment using GIZA++. On the one hand, if the reference translation (i.e. human translation) of the nouns is not equal $RT_1 \neq RT_2$, we do not want to translate consistently, and the instance is labelled as *None*. On the other hand, if both reference translations are consistent, we compare such translation RT with T_1 and T_2. If one of them is equal to the reference (i.e. $T_1 = RT$ or $T_2 = RT$), we then label the instance accordingly (e.g. if $T_1 = RT \neq T_2$, then $T_2 := T_1$). Finally, if none of them matches with the reference translation, we label it also as *None*.

We detect a total of 3,301 and 11,289 pairs in the UN Corpora for training the Chinese-English and German-English classifiers, respectively. The difference in the number of training distances is due to the amount of available data in each language pair: 10k sentences for Chinese-English versus 150k sentences for German-English. In this experiment, we keep very similar test set sizes. Specifically, 647 pairs on Chinese→English and 695 on German→English.

4.3.2 Classifiers for Consistent Translation

We train different classifiers on a set of syntactic, and semantic features to predict whether the system should use the same translation for a pair of repeated nouns, and if so, which of them is correctly translated. We describe in detail the semantic and syntactic features used to train the classifiers in section 4.3.2.1 and section 4.3.2.2, respectively.

To train the classifiers we use the WEKA environment,[6] which allows us to test several learning algorithms. Specifically, we train three different classifiers using different learning algorithms: C4.5 Decision Trees (J48 in Weka) (Quinlan, 1993), Random Forests (Breiman, 2001), and MaxEnt. For performance reasons, we train the latter with Stanford[7] instead of WEKA's Logistic Regression. To avoid overfitting, we use 10-fold cross validation on the training set.

We use the default settings from WEKA to train the classifiers. Specifically, for Decision Trees, the hyper-parameters *minNumObj* (i.e. number of instances per leaf) and the confidence factor for pruning are set to 2 and 0.25, respectively. We allow the Random Forests algorithm to generate 100 trees and set their maximal depth to unlimited. Finally, we set the tolerance of the MaxEnt algorithm, which is used for convergence in parameter optimisation to 10^{-5}.

In section 4.2, we integrate our method with the approaches `Decode` and `Post-editing` and conclude that the latter yields better results. First, the coverage of the `Decode` approach is lower than the `Post-editing`'s as in the former, we only enforce a translation if it is a translation candidate in the phrase table. Second, the `Decode` approach introduces translation candidates without probabilities, negatively affecting the quality of the overall sentence translation.

We therefore discard `Decode` in these experiments, and analyse the performance of `Post-editing` and `Re-ranking`, independently and combined. In the `Re-ranking` approach, we go through the first 10,000 translation hypotheses that the SMT system produces and select the first one that contains the translation of the noun pair as predicted by the classifier. If none of the hypotheses meets the condition, we keep the original best hypothesis proposed by the baseline. When we combine both `Post-editing` and `Re-ranking`, we automatically post-edit the translation in those cases where the translation is not found among the list of hypotheses.

[6] http://www.cs.waikato.ac.nz/ml/weka/
[7] http://nlp.stanford.edu/software/classifier.shtml

4.3.2.1 Semantic Features

The semantic features are divided into two groups: discourse and local context features, which differ in the amount of context they take into account. On the one hand, local context features represent the immediate context of each of the nouns in the pair and their translations. That is, three words to their left and three words to their right in both source and SMT output, always within the same sentence.

On the other hand, discourse features capture those cases where the inconsistent translations of a noun might be due to a disambiguation problem of the source noun, and semantic similarity can be leveraged to decide which of the two translations best matches the context. To compute the discourse features, we use the word vector representations generated from a large corpus using word2vec (Mikolov et al., 2013), which have been successfully used recently to compute the similarity between words (Schnabel et al., 2015). Specifically, we employ the model trained on the English Google News corpus with about 100 billion words.[8]

For each pair of inconsistent translations (T_1, T_2) of a source noun N, we compute the cosine similarities c_1 and c_2 between the vector representation of each translation and the mean vector of their contexts. These mean vectors, noted $\vec{v}_1$ and $\vec{v}_2$, are computed by averaging all vectors of the words in the respective contexts of T_1 and T_2. Here, the contexts consist of 20 words to the left and 20 words to the right of each T_i, possibly crossing sentence boundaries. The cosine similarities c_1 and c_2 are therefore:

$$c_1 = \cos(\vec{T_1}, \vec{v}_1) = \frac{\vec{T_1} \cdot \vec{v}_1}{||\vec{T_1}|| \cdot ||\vec{v}_1||}, \tag{4.1}$$

$$c_2 = \cos(\vec{T_2}, \vec{v}_2) = \frac{\vec{T_2} \cdot \vec{v}_2}{||\vec{T_2}|| \cdot ||\vec{v}_2||}. \tag{4.2}$$

The two values c_1 and c_2 are used as features, allowing classifiers to learn that, in principle, higher values indicate a better translation in the sense of its semantic similarity with the context.

In example 4.11, the German word *Politik* is translated into the English words *policy* and *politics*. The semantic similarity between the word *politics* and its context (c_2) is lower than the similarity between *policy* and its context (c_1), which we consider to be an

[8] https://code.google.com/p/word2vec

indication that the first occurrence, namely *policy*, has better chances to be the correct translation.

4.3.2.2 Syntactic Features

The syntactic features are defined under the assumption that the local parse tree of the noun occurrence that is correctly translated by the SMT system is syntactically more complex and therefore, has more local context bound to it. That is, for example, if N_1 (i.e. the first occurrence of a noun) is in a noun phrase (NP), where the number of words of the NP itself and its siblings is higher than in N_2, then N_1 has a higher probability of being correctly translated by the SMT system. This assumption is indeed confirmed by the obtained results.

These features are divided in three subsets that relate to the nouns themselves and their translations, the size of their siblings, and the size of their ancestors. In the first subset, we define (1) source noun, (2) distance in sentences between the two source occurrences, (3) translation of the first occurrence, and (4) translation of the second occurrence.

TABLE 4.9: Syntactic features and the corresponding values of the two occurrences of the German word *Politik* in example 4.13. Figure 4.1 shows the two parse trees of the sentences used to extract the values of the features.

Features	**Values**
Source Noun (German)	Politik
Distance in sentences between the two source occurrences	1
Translation of the 1^{st} occurrence	policy
Translation of the 2^{nd} occurrence	politics
Number of sibling nodes of the 1^{st} occurrence	2
Number of sibling nodes of the 2^{nd} occurrence	2
Sign of the difference between the above	0
Number of words of the 1^{st} occurrence and its siblings	1
Number of words of the 2^{nd} occurrence and its siblings	1
Sign of the difference between the above	0
Number of nodes in the first NP ancestor of the 1^{st} occurrence	12
Number of nodes in the first NP ancestor of the 2^{nd} occurrence	5
Sign of the difference between the above	1
Number of words in the first NP ancestor of the 1^{st} occurrence	5
Number of words in the first NP ancestor of the 2^{nd} occurrence	2
Sign of the difference between the above	1
Distance between the first NP ancestor and the 1^{st} occurrence	1
Distance between the first NP ancestor and the 2^{nd} occurrence	1
Sign of the difference between the above	0
Class (1, 2, 0)	1

```
(ROOT
  (S
    (VP (KOUI um) (VVIZU sicherzustellen) ($, ,)
      (S (KOUS dass)
        (NP (ART die) (ADJA Vereinten) (NN Nationen))
        (PP (APPR mit) (ART den) (ADJA raschen) (NN Entwicklungen)
          (PP (APPR in)
            (CNP
              (NP (ART der) (NN Informations))
              ($[ -) (KON und)
              (NP (NE Kommunikationstechnik) (NN Schritt)))))
        (VVFIN halten)))
    ($, ,) (VAFIN wurde)
    (NP (ART eine) (NN Politik)
      (NP (ART der) (ADJA Vereinten) (NN Nationen)))
    (VP
      (PP (APPR auf) (ART dem) (NN Gebiet)
        (NP (ART der) (NN Informationstechnik)))
      (VVPP erarbeitet))
    ($. .)))

(ROOT
  (S
    (PP (APPR unter) (NN Heranziehung)
      (NP (ADJA sekretariatsinterner) (NN Fachkenntnisse)))
    (VAFIN wird)
    (CNP
      (NP (PDAT diese) (NN Politik))
      (KON sowohl)
      (NP (ART die) (NN Einführung)
        (NP (KOKOM als)
          (CNP
            (NP (ADV auch) (ART die) (NN Handhabung)
              (NP (ADJA neuer) (NN Informationstechnologien)))
            (KON und)
            (NP (PRELAT deren) (NN Verwendung))))))
    (CVP
      (PP (APPR als) (NN Mittel))
      (PP (APPRART zur) (NN Informationsverbreitung))
      (KON und) ($[ -)
      (VP (ADV verwaltung) (VVINF bestimmen)))
    ($. .)))
```

FIGURE 4.1: Parse trees of the sentences in example 4.13 obtained with the Stanford parser. The blue boxes mark the analysed noun (i.e. *Politik*), and the red boxes correspond to the first NP ancestors.

We also compute the size of the siblings and ancestors in words and nodes. We therefore define the second subset as (1) number of sibling nodes of the first occurrence, (2) number of sibling nodes of the second occurrence, (3) sign of the difference between the number of sibling nodes of the first and second occurrence (+1, 0, -1), (4) number of words of the first occurrence and its siblings, (5) number of words of the second occurrence and its siblings, and (6) sign of the difference between the number of words of the first and second occurrence and their siblings (+1, 0, -1).

The remainder includes all features related to the ancestors such as (1) number of nodes in the first NP ancestor of the first occurrence, (2) number of nodes in the first NP ancestor of the second occurrence, (3) sign of the difference between the number of nodes in the

first NP ancestor of the first and second occurrence (+1, 0, -1), (4) number of words in the first NP ancestor of the first occurrence, (5) number of words in the first NP ancestor of the second occurrence, (6) sign of the difference between the number of words in the first NP ancestor of the first and second occurrence (+1, 0, -1), (7) distance between the first NP ancestor and the first occurrence, (8) distance between the first NP ancestor and the second occurrence, and (9) sign of the difference between the last two features (+1, 0, -1).

We exemplify the extraction of the syntactic features in the following. We first start with example 4.13, which contains the German noun *Politik* twice. While the SMT systems translates the first occurrence into *policy*, and the second into *politics*, the reference translates both into *policy*. We then obtain the parse tree of each sentence that are used to extract the syntactic feature values (See figure 4.1). Finally, table 4.9 shows the values of the features.

We observe in the obtained results that the first occurrence of *Politik* is correctly translated into *policy*, and has a higher number of nodes and words in the first NP ancestor than the second occurrence. This meets our assumption that the local parse tree of the noun occurrence that is syntactically more complex, has a higher probability of being correctly translated by the SMT system.

(4.13) **Source**: Um sicherzustellen, dass die Vereinten Nationen mit den raschen Entwicklungen in der Informations- und Kommunikationstechnik Schritt halten, wurde eine ***Politik*** der Vereinten Nationen auf dem Gebiet der Informationstechnik erarbeitet.
Unter Heranziehung sekretariatsinterner Fachkenntnisse wird diese ***Politik*** sowohl die Einführung als auch die Handhabung neuer Informationstechnologien und deren Verwendung als Mittel zur Informationsverbreitung und -verwaltung bestimmen.

SMT: To make sure that the United Nations with the raschen developments in the information and kommunikationstechnik step, was a ***policy*** of the United Nations, in the area of information technology out.
Under Heranziehung sekretariatsinterner Fachkenntnisse will this ***politics*** both the introduction, as well as the drive new information technologies, and their use as a means to determine Informationsverbreitung and -verwaltung.[9]

[9] We are aware that the translation contains many OOV words. Here, we just focus on the translation of the German *Politik*.

Human Reference: To ensure that the organization keeps up with the rapid developments in information and communication technology, a United Nations information technology ***policy*** has been developed.
Using in-house expertise, the ***policy*** will address both the introduction and management of new information technologies and their use as vehicles for the distribution and management of information.

4.3.3 Analysis of the Classification Task

The accuracy of the classification task, that is, the prediction of the correct translation variant (1^{st}, 2^{nd} or *None*) is above 80% on the development set (see table 4.10 and table 4.11) and 74-78% on the test set (see table 4.13 and table 4.12), for both language pairs. These are positive results as the number of instances per class is balanced. In these conditions, a baseline that makes a random prediction achieves only 33% accuracy.

The syntactic features outperform the semantic ones on the development set in both language pairs. When we combine both sets of features, we obtain the best results with the MaxEnt classifier in Chinese, and with random forests in German. In contrast, the other classifiers present lower scores. Overall, the MaxEnt classifier performs best on the test set in both language pairs, and with the three sets of features (i.e. syntactic, semantic, and the combination of both) (see figure 4.2), followed by Random Forests. Even though Random Forests performs best on the development set in German, the accuracy of the MaxEnt classifier is only slightly lower. We therefore conclude that the best configuration to train our classifiers is the MaxEnt classifier with all features.

The BLEU scores show that `Re-ranking` and `Post-editing` together outperform their individual application for all features and both language pairs. The scores rise from 11.07 to 11.36 in Chinese→English and from 17.10 to 17.67 in German→English.

TABLE 4.10: Results on the prediction of the class label (1, 2, or *None*) for repeated nouns in **Chinese**, in terms of accuracy (%) and *kappa* scores on the development set with 10-fold cross-validation. Methods are sorted by average accuracy over the 3 feature sets.

	Syntactic features		Semantic features		All features	
	Acc. (%)	κ	Acc. (%)	κ	Acc. (%)	κ
J48	72.1	0.48	60.2	0.00	60.2	0.00
RF	75.3	0.54	68.4	0.29	70.7	0.35
MaxEnt	76.7	0.65	69.5	0.32	**83.3**	**0.75**

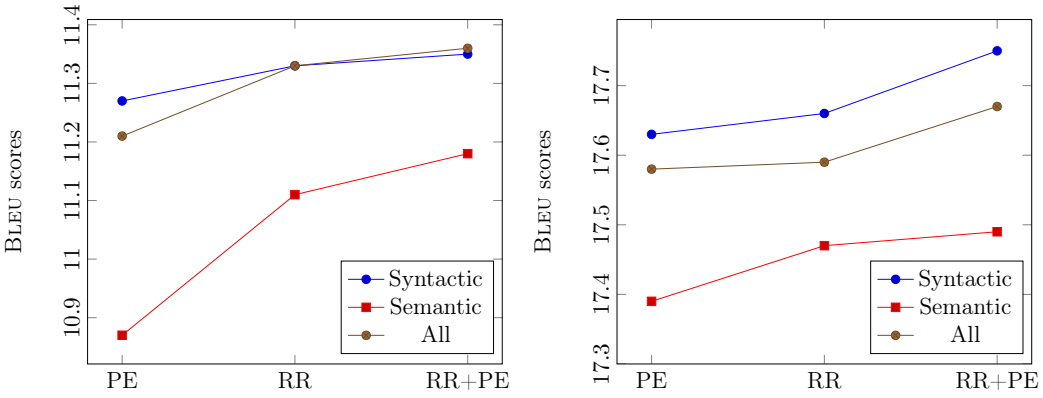


FIGURE 4.2: Translation quality (BLEU) of the Maximum Entropy classifier on the Chinese-English test set (left) and German-English test set (right). The figures show the translation quality of the three proposed approaches (`Post-editing` (PE), `Re-ranking` (RR), and a combination of both (RR+PE)) on the test sets using syntactic, semantic, or syntactic and semantic features.

Note that the largest improvement for this test set in German→English is achieved when only syntactic features are used (i.e. 17.75). The oracle BLEU scores shows us the scores that our method would achieve with ideal results. We observe that our method covers 51% of the BLEU gap between the baseline and the oracle systems on Chinese→English and 64% on German→English.

To get a better insight on the low performance of the semantic features, we inspect the test sets manually and observe that the contexts of the noun translations are similar. Table 4.14 shows an analysis of the effect of the semantic features on different training sets in terms of accuracy and kappa scores.

These training sets are built according to the cosine similarity between T_1 and T_2, as follows: for each training instance (pair of nouns), we compute the cosine similarity between the vector representation of T_1 and T_2. Next, we group instances by intervals and carry out 10-fold cross validation for each subset. The lower the range values, the more dissimilar the translation pairs T_1 and T_2, and the better the scores of discourse features. Specifically, when the translations are dissimilar, the classifier makes better

TABLE 4.11: Results on the prediction of the class label (1, 2, or *None*) for repeated nouns in **German**, in terms of accuracy (%) and *kappa* scores on the development set with 10-fold cross-validation. Methods are sorted by average accuracy over the 3 feature sets.

	Syntactic features		Semantic features		All features	
	Acc. (%)	κ	Acc. (%)	κ	Acc. (%)	κ
J48	77.0	0.66	64.8	0.45	79.7	0.69
RF	82.0	0.73	73.5	0.60	84.5	0.77
MaxEnt	80.8	0.71	76.8	0.65	**83.4**	**0.75**

TABLE 4.12: Prediction of the correct translation (accuracy (%) and *kappa*) and translation quality (BLEU) for repeated nouns on the **Chinese test set**. Maximum Entropy was the best method found on the dev set.

	Syntactic features					Semantic features					All features				
	Acc.	κ	BLEU			Acc.	κ	BLEU			Acc.	κ	BLEU		
			PE	RR	RR+PE			PE	RR	RR+PE			PE	RR	RR+PE
Baseline	-	-	11.07	11.07	11.07	-	-	11.07	11.07	11.07	-	-	11.07	11.07	11.07
J48	66.3	0.42	11.17	11.20	11.30	33.1	0.00	11.07	11.07	11.07	33.1	0.00	11.07	11.07	11.07
RF	71.7	0.53	11.22	11.24	11.27	55.2	0.33	11.04	11.07	11.12	54.9	0.32	11.16	11.20	11.24
MaxEnt	73.7	0.60	11.27	11.33	**11.35**	56.1	0.34	10.87	11.11	11.18	72.5	0.56	11.21	11.33	**11.36**
Oracle	100	1.00	11.40	11.52	11.64	100	1.00	11.40	11.52	11.64	100	1.00	11.40	11.52	11.64

TABLE 4.13: Prediction of the correct translation (accuracy (%) and *kappa*) and translation quality (BLEU) for repeated nouns on the **German test set**. Maximum Entropy was the best method found on the dev set.

	Syntactic features					Semantic features					All features				
	Acc.	κ	BLEU			Acc.	κ	BLEU			Acc.	κ	BLEU		
			PE	RR	RR+PE			PE	RR	RR+PE			PE	RR	RR+PE
Baseline	-	-	17.10	17.10	17.10	-	-	17.10	17.10	17.10	-	-	17.10	17.10	17.10
J48	70.5	0.56	17.59	17.61	17.70	48.2	0.23	17.13	17.27	17.33	69.4	0.54	17.56	17.60	17.66
RF	70.2	0.55	17.55	17.62	17.68	54.4	0.32	17.21	17.34	17.37	67.6	0.52	17.53	17.57	17.63
MaxEnt	78.3	0.67	17.63	17.66	**17.75**	63.5	0.49	17.39	17.47	17,49	68.7	0.53	17.58	17.59	**17.67**
Oracle	100	1.00	17.78	17.83	17.99	100	1.00	17.78	17.83	17.99	100	1.00	17.78	17.83	17.99

predictions with the discourse features (i.e. considering a larger context). However, the more similar the words are, the better the local context features (i.e. the surrounding words).

In example 4.14 we deal with the German word *Absatz* that is inconsistently translated into *paragraph* and *heels*. These translations are not interchangeable as they are used in different contexts. Therefore, the discourse features (i.e. the cosine similarities between the vector representation of each translation and the mean vector of their contexts.) of *heel* and *paragraph* are 0.11 and 0.35, respectively. This difference of 0.24 tells the classifier that *paragraph* fits better in the context than *heel*.

(4.14) **Source**: Ist dem Vertragsstaat, der seine Gerichtsbarkeit nach ***Absatz*** 1 oder 2 ausübt, mitgeteilt worden oder hat er auf andere Weise Kenntnis davon erhalten, dass einer oder mehrere andere Vertragsstaaten in Bezug auf dasselbe Verhalten Ermittlungen, Strafverfolgungen oder ein Gerichtsverfahren durchführen, setzen sich die zuständigen Behörden dieser Vertragsstaaten gegebenenfalls miteinander ins Benehmen, um ihre Maßnahmen abzustimmen.
Dieser Artikel findet Anwendung auf die Straftaten nach diesem Übereinkommen oder in Fällen, in denen eine organisierte kriminelle Gruppe an einer in Artikel 3 ***Absatz*** 1 Buchstabe a oder b. . .

SMT: Is the Vertragsstaat, the ***heel*** of his Gerichtsbarkeit after one or two wield shared been or did it in a different way of knowledge, that one or several other Vertragsstaaten in terms of the same behavior investigation, Strafverfolgungen or a trial, put the authorities zuständigen this Vertragsstaaten review together into the behavior, their actions to vote.
This article finds application on the crimes after this arrangements or in cases where an organized criminal group at an article in three ***paragraph*** one letter a or b. . .

TABLE 4.14: Effects of semantic similarity on classification (10-fold cross validation). The scores with discourse features increase as similarity between T_1 and T_2 decreases.

		Local Context		Discourse		Both	
cosSim.	Inst.	Acc.	κ	Acc.	κ	Acc.	κ
0.0–0.1	141	63.8	0.27	73.8	0.47	66.0	0.31
0.1–0.2	341	70.1	0.40	75.4	0.51	71.0	0.42
0.2–0.3	350	73.1	0.43	68.0	0.35	72.3	0.41
0.3–0.4	350	72.6	0.45	66.0	0.32	68.6	0.37

Human Reference: If a State Party exercising its jurisdiction under ***paragraph*** 1 or 2 of this article has been notified, or has otherwise learned, that one or more other States Parties are conducting an investigation, prosecution or judicial proceeding in respect of the same conduct, the competent authorities of those States Parties shall, as appropriate, consult one another with a view to coordinating their actions.
This article shall apply to the offences covered by this Convention or in cases where an offence referred to in article 3, ***paragraph*** 1 (a) or (b)...

In contrast, when the translations can be found in similar contexts, their values of the discourse features are also similar. In example 4.15, the discourse features of *measures*, and *action* are 0.29, and 0.25, respectively. In these cases, the classifier makes better predictions with the local context features.

(4.15) **Source**: Er besteht darauf, dass die Taliban aufhören, internationalen Terroristen und ihren Organisationen Zuflucht und Ausbildung zu gewähren, dass sie wirksame ***Maßnahmen*** ergreifen, um sicherzustellen, dass das unter ihrer Kontrolle befindliche Gebiet nicht für terroristische Einrichtungen... Der Sicherheitsrat wird die wirksame Durchführung der mit dieser Resolution auferlegten ***Maßnahmen*** sicherstellen.

SMT: He insists that the Taliban stop, international terrorists and their organizations refuge and education to allow you to effective ***measures***, to make sure that the under your control situated area not for terrorist institutions...
The security council is the effective do with this resolution auferlegten ***action*** to make sure.

Human Reference: It insists that the Taliban cease the provision of sanctuary and training for international terrorists and their organizations, take effective ***measures*** to ensure that the territory under its control is not used for terrorist installations...
The Council will ensure effective implementation of the ***measures*** imposed by that resolution.

Despite the difficulty of semantic features at predicting the correct translation, we observe that they perform better in German than in Chinese. As Huang (1995) states, strong polysemy or homonymy is less frequent in Chinese than in English. We hypothesise that this statement extends to German, and therefore, German texts are more likely to contain nouns whose translations into English are semantically divergent.

We rank the features by information gain using Weka, and observe that, in both language pairs, the ranking is headed by the features: source noun, translation of the first, and translation of the second occurrence, which provide information rather lexical than syntactic. The rank continues with features in the subset that relates to the size of the ancestors, which are more syntactic.

4.4 Summary

In this chapter, we analysed consistency in translation under specific conditions: references to compounds, and pairs of repeated nouns in a document. In the first case, our method encourages a consistent translation of the references using the translation of the nominal head of the compound. We take advantage of compounds as they have less translation variants than words with only one morpheme, and therefore, they are less prone to ambiguity. We then extended the issue to repeated nouns. We trained several classifiers to predict if there is a correct translation of a word among all its occurrences in the document and, if so, which of the translations we should consistently use.

Our experiments showed that consistency on the translation leads to a slight improvement of the quality of the translations. Note that we did not tackle the consistent translations of all words in a document, and therefore, the margin of ideal improvement shown by the oracle translations was relatively small.

The main reason to tackle only these two specific cases was to avoid too much repetition that would affect fluency. However, it is still unclear how much consistency is enough or too much. What we certainly know is that there are no *perfect* synonyms, that is, words that share exactly the same meaning, or at least, they are extremely rare (Lyons, 1968). It would be then a great challenge to encourage lexical variability in Statistical Machine Translation, as it is to find perfect synonyms in the context of the document. Therefore, finding the right translation and applying it consistently is a better strategy for MT to ensure a correct translation of that word in the document. We believe that there are cases where consistency is expected, as in the cases of reference tackled in this chapter, that is, either by using the nominal head of the compound or the repetition of a noun.

We believe that in a perfect scenario, consistency would be encouraged only in those translations that present an ambiguity issue. For example, we found human translations of the German noun *Wand* in French, such as *face*, *paroi*, and less frequently *mur*. All these translations can refer to the sense of "the wall of a mountain", and therefore, it would be completely fine to alternate them in the majority of cases. However, the translations of polysemic words such as *Absatz* ("heels", "paragraph", or "sales") cannot

be interchanged. It is then very important to be able to find the right translation in those cases and apply it consistently.

In the next chapter, we extend our coverage, considering not only nouns, but also other content words, such as adjectives and verbs. In addition, we broaden the discourse context to improve the lexical choice. Specifically, while in this chapter we only rely on the translation of a compound or another occurrence of the same word, in the next experiments, we take into account semantically-similar words from the discourse.

Chapter 5

Exploiting Lexical Chains in SMT

Lexical chains are chains of semantically related words, which represent the structure of a document (section 5.1). In this chapter, we present a method that utilises the context provided by the lexical chains to improve the translation output of SMT. Specifically, the method improves the lexical choice of words in the chain that have multiple translations and are ambiguous in the context of the sentence.

The method consists of first detecting the lexical chains on the source side and then, keeping the semantic similarity of words in the counterpart target chains. In order to use the lexical chain detected from the source to the target, we implement and integrate a feature function into the document-level decoder Docent in section 5.2 (see section 1.3.1.2 for a technical overview of the decoder).

To assess the performance of the presented method, we carry out several experiments on the translation from German into English (section 5.4). One of the key ingredients of the method is that it uses word embeddings to detect the lexical chains instead of external lexical resources (section 5.2.3). When using word embeddings, for example, it is possible to handle words other than nouns much more easily than with lexical resources that focus mostly on nouns. The translation output of our method is compared to a baseline that does not handle lexical chains and also to a method that uses external lexical resources. The manual evaluation shows that the presented method improves 36%-48% of the translation changes over the baseline (section 5.4).

Finally, we carry out a study on the three parameters that define the strength of a lexical chain: length, density and repetition (section 5.5). Our study aims at finding out the relevance of each of these parameters when computing the strength of a lexical chain. To do so, we set them to different weights and assess the impact on the translation output.

5.1 Introduction to Cohesion and Lexical Chains

In chapter 1, we introduce the issues of state-of-the-art phrase-based SMT systems that deal with sentences in isolation. Documents are a set of sentences that function as a unit. When we translate at document-level we take into account document properties that help to improve the quality of the translation, not only locally, but also in the context of the document.

Coherence and cohesion are terms that describe properties of texts. On the one hand, coherence concerns the semantic meaningfulness of the text. On the other hand, cohesion has to do with relating the sentences throughout the text, which is achieved through reference, ellipsis, substitution, conjunction, and the use of semantically-similar words. Often, semantically-similar words are related sequentially in a text, defining the topic of the text segment that they cover. These sequences of words are called lexical chains.

(5.1) Die letzte Laureatin vergab den Nobelpreis für Ökonomie, den die Amerikanerin Elinor Ostrom und ihr Landsmann Oliver Williamson für Analysen des Wirtschaftsberichtes erhalten haben.
Die einzige Fachkategorie, in der in diesem Jahr keine Frau einen Nobelpreis erhielt, war Physik.
Diesen Preis haben heute die Wissenschaftler Charles Kao für die Forschung im Glasfasern-Bereich . . . davongetragen . . .
Jeder dieser Preisträger hat das Diplom, die Nobelmedaile und Bescheinigung über den Erwerb des Geldpreises erhalten.

Lexical Chain 1: {*Laureatin* ("Laureatin"), *vergab* ("awarded"), *Nobelpreis* ("Nobel Prize"), *Nobelpreis* ("Nobel Prize"), *Preis* ("prize"), *Preisträger* ("prize winner"), *Diplom* ("diploma"), *Nobelmedaile* ("Nobel medal"), *Bescheinigung* ("Certificate"), *Geldpreises* ("monetary prize")}

Lexical Chain 2: {*Ökonomie* ("Economics"), *Analysen* ("analysis"), *Wirtschaftsberichtes* ("economic reports"), *Physik* ("Physics"), *Wissenschaftler* ("scientists"), *Forschung* ("research")}

The example above (5.1) shows two possible lexical chains that we manually extracted from five sentences. The text fragment of a document from newstest2010.[1] When we

[1] http://www.statmt.org/wmt16/translation-task.html

look at the words that comprise each lexical chain, we observe that one of them is about *awards* and the other about *science*. Indeed, lexical chains define the meaning of the text segment that they cover.

In this chapter, we assume that semantic relatedness of the words that constitute the lexical chains in the source document must be preserved in the translation. We therefore focus on lexical chains as a means to keep the semantic similarity from the source to the target. The idea is to utilise the related words in the same lexical chain to get a better context of every word in the chain and improve lexical choice in the translation. Without this discourse knowledge, the decoder generally produces wrong translations, when words are ambiguous in the context of the sentence and the correct translation is not the most frequent in the training data.

Consider again the example (5.1). The German word *Preis* in *Lexical Chain 1* has two senses: *price* or *award*, which results in two different translations in English at least. When any document-level information is ignored, *Preis* is mistranslated into *price*, breaking the semantic similarity between the words in the target lexical chain. Indeed, the lexical chain in the target does not longer keep the real meaning of this fragment, since it contains a word in the wrong sense. If we take advantage of lexical chains to improve translation, we then get the context from the words *Nobelpreis* ("Nobel Prize") and *Preisträger* ("prize winner") that are linked to *Preis* in the lexical chain and produce the right translation.

Note that *Preis* is a reference of the previous occurrences of the compound *Nobelpreis*. In this case, the right translation could be also produced by looking at the translation of the compound, as discussed in chapter 4. Lexical chains are therefore an extension of the consistency problem, since they consist of any semantically-similar word and are used to improve the translation of any of their words.

Lexical chains have been successfully used in other research areas such as information retrieval (Stairmand, 1996, Rinaldi, 2009) and document summarization (Barzilay and Elhadad, 1997, Pourvali and Abadeh, 2012), but they have received little attention in MT.

Galley and McKeown (2003) introduce a method to detect lexical chains using WordNet (Miller, 1995). The method first builds a representation of all words in the document and all their senses, creating semantic links such as synonym, hypernym, hyponym, and sibling between them. It then uses the semantic links to disambiguate each word and builds the lexical chains accordingly.

Galley and McKeown (2003) evaluate the performance of the method on a sense disambiguation task. Indeed, lexical chains help to disambiguate the sense of polysemic words

by looking at the words in the chain. Despite of the problems of word senses (Kilgarriff, 1997, 2006, Hanks, 2000), it shows the capability that lexical chains have to improve the lexical choice of words with multiple translations in MT. As introduced in section 2.1, Xiong et al. (2013b) are the first to explore the benefits of using lexical chains in Statistical MT from Chinese into English, using Galley and McKeown (2003)'s method to detect the lexical chains in the source.

In this chapter, we present a simpler method as it does not use external lexical resources and builds classifiers per each word. Instead, it uses word embeddings to detect the lexical chains in the source and also to maintain the semantic similarity of the detected lexical chains on the target side. The method is integrated into the document-level SMT decoder Docent, and we report experimental results on the translation from German into English.

5.2 A Lexical Chain Model for SMT

This section describes our method of improving the quality of translation in Statistical Machine Translation utilising lexical chains. The method works as follows: it first detects the lexical chains in the source document[2] (section 5.2.1) and feeds them into the Lexical Chain Translation Model (LCTM), which is integrated into the document-level decoder Docent.[3] The model then gets their counterpart in the target through word alignment and computes the LCTM score that contributes to the overall translation score in the Statistical MT system (section 5.2.2).

5.2.1 Building Source Lexical Chains

Our method to detect and build lexical chains from a document is inspired by the approach proposed by Morris and Hirst (1991). Their approach consists of manually detecting those lexical chains by applying a thesaurus to find the similarity between words. Our method implements the manual algorithm, detecting and building the lexical chains automatically.

Instead of applying an external lexical resource, we apply word embeddings to compute the semantic similarity. Word embeddings are representations of words in a vector space

[2] A version of the code to detect lexical chains is available at https://github.com/lmascarell/lexCH. This version uses the SenseGram tool (https://github.com/tudarmstadt-lt/sensegram) to obtain sense embeddings.

[3] The code of Docent and the LCTM is available at https://github.com/lmascarell/docent

and are commonly exploited to compute similarity between words (Mikolov et al., 2013) (See discussion in section 5.2.3).

The method works as follows. It processes sentences in a given document sequentially. For each content word c in every sentence, it checks whether c is semantically related to the previous content words c' in a span of five sentences, as suggested by Morris and Hirst (1991). If c and c' are semantically related, we proceed as follows:

- If c and c' do not belong to any chain, we create a new chain consisting of c and c'.
- If c' is in a chain ch_i, we append c to ch_i.
- If c and c' belong to two different chains, we then merge both chains.

The detected lexical chains preserve the semantic link between related content words, creating also one-transitive links. That is, c_i links to c_{i+l} by transitivity if c_i links to c_{i+k} and c_{i+k} to c_{i+l}, where $i<k<l$ (Morris and Hirst, 1991).

Every link to a word in the lexical chain gives context to disambiguate the word itself. Therefore, the more links are created, the better. One-transitive links are safe to consider, because they are still semantically related, as indicated by Morris and Hirst (1991), but further than that leads to errors. Morris and Hirst (1991) point to the following lexical chain in their paper: {*cow*, *sheep*, *wool*, *scarf*, *boots*, *hat*, *snow*}. Here, we observe that while consecutive words in the chain like *wool* and *scarf* are semantically related, *cow* and *snow* are not.

Example 5.2 shows the output of our method on this fragment of text, where each colour corresponds to a different lexical Chain. The same output is illustrated in figure 5.1, which shows two graphs that correspond to the two lexical chains.

(5.2) Ihr nächstes *Smartphone* wird zwei *Betriebssysteme* beherrschen.
Die Amerikaner rechnen für die Zukunft mit einem *Handy*, auf dem der *Benutzer* durch drücken einer einzigen Taste zwischen verschiedenen *Betriebssystemen* *umschalten* kann.
Die vorgelegten Pläne sehen vielversprechend aus.

Lexical Chain 1: {*umschalten* (“switch”), *Betriebssystemen* (“operating system”), *Benutzer* (“user”), *Betriebssysteme* (“operating system”)}

Lexical Chain 2: {*Handy* (“cell phone”), *Smartphone* (“smart phone”)}

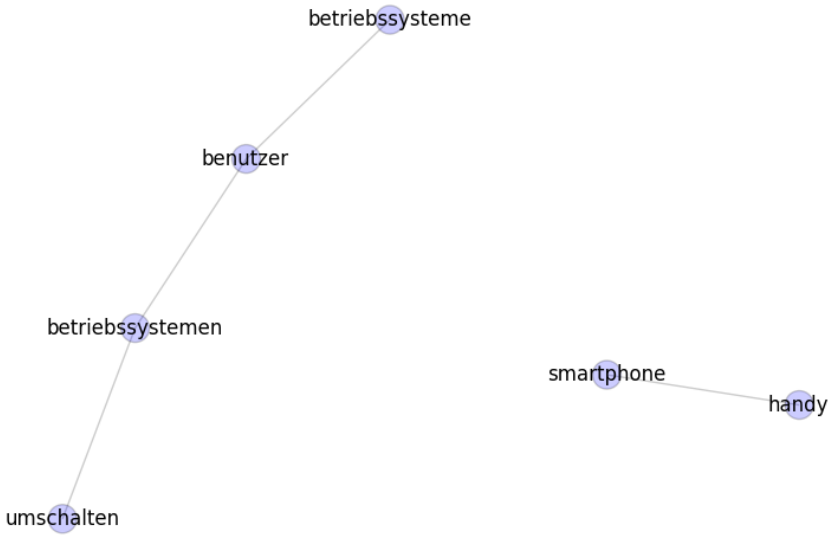


FIGURE 5.1: Lexical chains detected with the method on the three sentences from example 5.2.

5.2.2 The Lexical Chain Translation Model

In order to improve translation quality utilising lexical chains, we develop a model that favours document translations where the words in the target lexical chain are semantically related. The target lexical chains are the corresponding counterpart of the source lexical chains detected, which are obtained by the Lexical Chain Translation Model (LCTM) through word alignment. The LCTM also uses word embeddings to compute the semantic similarity between words and it is integrated as an additional feature function in the document-level decoder Docent as a standard Statistical Machine Translation model (equation 1.1)

To understand how the model is integrated into Docent, it is important to understand how Docent works (see details in section 1.3.1.2). In a nutshell, Docent implements a search procedure based on local search. At every stage of the search, the decoder produces a random change in the document translation. The search algorithm accepts a new state (i.e. a new translation of the document), when its document score computed by equation 1.1 is higher than the last accepted. To compute the document score, it considers the score obtained from each feature function. The initial translation of the whole document is either randomly generated or a translation from Moses.

The LCTM is implemented as one of the feature functions in Docent, and therefore, it contributes to the overall document score. Consider the example 5.3. This example shows

two possible translations of the sentence in example 5.1 *Diesen Preis haben heute die Wissenschaftler Charles Kao für die Forschung im Glasfasern-Bereich ... davongetragen ...*, which belong to a different Docent state. Since *Preis* is linked to *Nobelpreis* ("Nobel Prize") and *Preisträger* ("prize winner") in the source lexical chain, the semantic similarity of its counterpart lexical chain in the target is higher when *Preis* is translated into *award*. This leads to a higher LCTM score that contributes to a higher document score. The state q is then preferred by the decoder. Note that in this case, the language model also increases in state q. That is because *received* has a higher probability together with *award* than with *price*.

(5.3) a. State q: This *award* was received today by scientists Charles Kao for research in the field of optical fibers ...

b. State r: This *price* was received today by scientists Charles Kao for research in the field of optical fibers ...

Lexical Chain: {..., *Nobelpreis*, *Preis*, *Preisträger*, ...}

Computation of the Model Score

Each lexical chain is a chain of words connected by their semantic similarity. We define the model score as the mean of the semantic relatedness scores of each target lexical chain in a document translation. To compute the semantic relatedness sim_i of a lexical chain ch_i, we average the semantic similarity of all links in ch_i as in the following equation

$$sim_i = \frac{1}{m} \sum_{j=1}^{m} SemLink_{ij}, \tag{5.1}$$

where every link is comprised of two words and its semantic similarity *SemLink* is the cosine distance between their embeddings. In the experiments, we use German in the source, which is a language rich in compounds. These compounds have multiword equivalents in English and can be detected as part of a lexical chain (e.g. *Nordwand* is translated into the English *north face*). To deal with such cases, sim_i is the maximum similarity score obtained from each content word in the translation of a compound and the rest of the words in the lexical chain.

Every lexical chain has a different *relevance* in the computation of the LCTM score, which depends on three factors introduced by Morris and Hirst (1991): length (λ), repetition (β), and density (ρ). The later is defined as the ratio of words in the lexical chain to all words in the fragment of text that it covers. Accordingly, the longer, the denser the lexical chain is and the more repetition it has, the higher its weight is in the computation

of the overall model score. These factors have not been addressed in the literature when dealing with lexical chains. Morris and Hirst (1991) define the *strength* of lexical chains (i.e. *relevance*), but they do not use it in their experiments.

To compute the length, density, and repetition of every lexical chain (i.e. λ_{ch_i}, ρ_{ch_i} and β_{ch_i}) we proceed as follows. Let *rel* be the total number of semantic relations in a lexical chain ch_i, *rep* the total number of repetitions, and *span* the number of words in the fragment of the document between the head and the tail of ch_i. ρ_{ch_i} and β_{ch_i} are then computed by the following two equations

$$\rho_{ch_i} = \frac{rel}{span}, \tag{5.2}$$

$$\beta_{ch_i} = \frac{rep}{span}. \tag{5.3}$$

Finally, the length λ_{ch_i} is the ratio of *rel* to the number of relations of the longest lexical chain detected. The longest lexical chain gets therefore the highest length value (i.e. 1.0) among all lexical chains in the document.

After computing all factor values for each lexical chain, the model computes the weight for each of them. The weight w of a chain ch_i is then the average of ρ_{ch_i}, λ_{ch_i} and β_{ch_i}, where ρ_{ch_i}, λ_{ch_i}, β_{ch_i}, and w_{ch_i} are all values between 0 and 1.

Finally, the overall LCTM score is computed by

$$LCTM = \frac{1}{n}\sum_{i=1}^{n} w_{chi} \cdot \frac{1}{m_i}\sum_{j=1}^{m_i} SemLink_{ij}. \tag{5.4}$$

5.2.3 Computation of Semantic Similarity

Dictionaries have been described in the literature to deal not only with lexical chains (Galley and McKeown, 2003), but with any task related to semantics such as Word Sense Disambiguation (WSD). However, it is unrealistic to assume that the fine-grained classification of senses in dictionaries is adequate for any NLP application (Kilgarriff, 2006). Even the classification itself has been questioned in terms of cognitive validity (Kilgarriff, 1997, 2006, Hanks, 2000).

The method presented in this chapter uses word embeddings as a means to compute semantic relatedness between words independently of dictionary senses (Mikolov et al., 2013). They have been indeed recently proposed for WSD tasks (Iacobacci et al., 2016).

Previously, other approaches were introduced to utilise embeddings for supervised (Zhong and Ng, 2010, Rothe and Schütze, 2015, Taghipour and Ng, 2015) and knowledge-based WSD (Chen et al., 2014).

As Firth (1957) stated "You shall know a word by the company it keeps." That is, words that are used and occur in the same contexts tend to have similar meanings. Essentially, word embeddings are vector representations of words in a vector space that are learned based on the immediate context in which they occur.

The coverage and the quality of the lexical chains are the most important factors in our approach to improve translation. Words that are not in any lexical chain are not considered for improvement at the decoding stage by our LCTM. Word embeddings detect words as semantically related when they occur in similar context, even if they do not have a hypernym, hyponym or sibling relation. Halliday and Hasan (2014) define the words that do not have a traditional sense relation, but belong to the notion of lexical cohesion as *collocations*. The lexical chain detection method includes them in the same lexical chain, since they also help to disambiguate the translation of a word. For example, the word *climber* can be related to *mountain* with word embeddings, but not with Galley and McKeown (2003)'s approach.

The main problem of word embeddings arises from words with multiple senses that are not disambiguated in the training phase. That is, each word has only one vector representation, including those polysemic words. For example, consider the English word *play*, which appears in different contexts such as to perform on a musical instrument, to take part in a sport or game, and to interpret a role. The word embedding then represents all senses together. Consequently, the semantic similarity between *play* and *guitar* is low, because the similarity is computed between *guitar* and all the senses of *play* together.

Word senses need to be disambiguated in the training phase to generate distinct vector representations for each sense. We therefore employ a method introduced by Thater et al. (2011), which uses the syntactic information to build *contextualized* embeddings. Consider again the word *play*, which appears in the sentences *I play the piano*, *we play the guitar*, *we play tennis*, *they play football*, and *they play Hamlet*. The approach proposed by Thater et al. (2011) extracts all the syntactic relations such as subject or object, resulting in three contextualised vectors built upon (1) *I play the piano*, *we play the guitar*; (2) *we play tennis*, *they play football*; and (3) *they play Hamlet*. The approach groups sentences in the same context by computing the semantic similarity between the context words (e.g. *piano* and *guitar*).

Finally, to compute the semantic similarity of two words, the proposed method computes the cosine similarity between their vector representation $\vec{a}$ and $\vec{b}$ as follows:

$$\cos(\vec{a}, \vec{b}) = \frac{\vec{a} \cdot \vec{b}}{||\vec{a}|| \cdot ||\vec{b}||}. \tag{5.5}$$

The closer to 1.0 the resulting value is, the more similar they are. We set a threshold of 0.45 to distinguish between similar and dissimilar words. This threshold is manually picked by looking at how different values impact on the resulting lexical chains. A lower threshold introduces too many words that are mostly related by their part-of-speech. A higher threshold results in semantically strong lexical chains, but misses out on words that are also related.

5.3 Setup of the Lexical Chain Translation Task

We conducted several experiments to prove the efficacy of the lexical chain detection and Lexical Chain Translation Model (LCTM) in Statistical MT. Lexical chains are difficult to evaluate in isolation, and therefore their quality is usually evaluated on the basis of the application for which they are used. Thus, we assess the performance of the method on the German→English translation task. We then compare it to the algorithm presented by Galley and McKeown (2003), which uses external resources instead of word embeddings to build the chains.

To build the lexical chains following Galley and McKeown (2003)'s method, we use GermaNet (Hamp and Feldweg, 1997) as external resource on the German side. The detected lexical chains are automatically annotated in the MMAX format[4] and then fed into Docent.

We also evaluate the performance of our LCTM compared to a variation that gives the same weight to all lexical chains, independently of their length, density or the repetitions. This new model, which we call LCTM$_{base}$, allows us to assess the importance of considering these three factors in the computation of the LCTM score.

We build a German→English phrase-based SMT system using data from Europarl and News Commentary in equal parts (see table 5.1) with the standard training settings described in section 1.3.4. The rest of the data used for tuning and testing comes from

[4] http://mmax2.sourceforge.net

TABLE 5.1: Total of segments per language pair from Europarl and News Commentary used to train the German→English phrase-based SMT system.

	Training	Tuning	LM
Lines	400K	5K	570K
Tokens	∼ 11M	∼ 125K	∼ 15M

the WMT'16 translation task. Specifically, we use the first 17 documents of newstest2010 (375 segments) as a development set of the LCTM and LCTM$_{base}$ and newstest2011 (554 segments), newstest2012 (684 segments), and newstest2013 (1,053 segments) for testing. We refer to section 1.3.5 for a more detailed description of the data.

The method uses word embeddings to detect the source lexical chains. We therefore train a skip-gram 300-dimensional model in German using the word2vec tool.[5] The texts come mainly from SdeWaC (Faaß and Eckart, 2013) (∼768M words)[6] and Common Crawls (∼775M words),[7] which are monolingual corpora collected from web sources. The rest of the data is from Europarl (∼47M words) and News Commentary (∼6M words). The Lexical Chain Translation Model (LCTM) model also needs to compute the similarity of the words in the target lexical chains. For this purpose, we use a skip-gram 300-dimensional model trained on English Google News (∼100 billion words).[5]

5.4 Experimental Results

In this section, we present the results obtained through the combination of lexical chain detection (using word embeddings and GermaNet) and the Lexical Chain Translation Model (LCTM). The LCTM takes into account the relevance (i.e. strength) of every lexical chain to compute the overall score. We then perform a third experiment that ignores this fact to assess its impact in the translation quality. To do so, we develop a model that behaves like the LCTM, except that it assigns the maximum strength value (i.e. 1.0) to all lexical chains. We refer to this new model in the following as LCTM$_{base}$.

The results of the experiments show between 20 to 30 translation changes in every test set due to lexical chains. We observe that the translation changes are often correct although it does not use the same terms as in the reference. Therefore, the fluctuations in BLEU scores are small (± 0.1), and so BLEU does not provide sufficient insight into the performance.

[5] https://code.google.com/p/word2vec
[6] http://www.ims.uni-stuttgart.de/forschung/ressourcen/korpora/sdewac.en.html
[7] http://www.statmt.org/wmt16/translation-task.html

TABLE 5.2: Manual evaluation results of the presented method (1) compared to using GermaNet for lexical chain detection (2). The analysis shows the percentage of correct (+) and wrong translations (-), and the improvement over the baseline (++). There are a total of 20 to 30 translation changes in every test set due to the lexical chains. We observe that the method (1) outperforms the approach that uses GermaNet (2). It also performs better than the method that ignores length, density, and repetition for the computation of the strength of each lexical chain in the overall score (3).

	newstest2011			newstest2012			newstest2013		
	+	−	++	+	−	++	+	−	++
Word Emb. & LCTM (1)	0.81	0.19	0.48	0.88	0.12	0.36	0.83	0.17	0.39
GermaNet & LCTM (2)	0.71	0.29	0.38	0.62	0.38	0.31	0.65	0.35	0.35
Word Emb. & LCTM$_{base}$ (3)	0.64	0.36	0.22	0.67	0.33	0.18	0.61	0.39	0.16

We then perform a manual evaluation to assess the results of the experiments. The annotation is carried out by two annotators who judge the quality of the translation changes due to the lexical chains. Specifically, the annotators obtain for each translation change the source sentence, the baseline (i.e. the translation ignoring lexical chains), the translation produced by the method we want to evaluate, and the reference. They then annotate whether the word that changes due to lexical chains is better than the one produced by the baseline, equally good or worse. For instance, consider the following annotation example extracted from newstest2012:

(5.4) **Input** Keine befreiende Novelle für Tymoshenko durch das Parlament
Baseline No liberating novella for Tymoshenko by Parliament
Lexical Chains No liberating amendment for Tymoshenko by Parliament
Human Reference Parliament does not support amendment freeing Tymoshenko

The German noun *Novelle* can be translated either into the English *novella* or *amendment*, which are in the sense of *narration* and *law*, respectively. In this example, the correct sense is picked by the method that exploits lexical chains, since *Novelle* is linked in the lexical chain to *Neuregelung* ("new regulation"). The translation is therefore annotated as better than the baseline.

The inter-rater agreement measured in terms of Cohen's Kappa score (Cohen, 1960) is 0.78 for the experiment that uses GermaNet to detect lexical chains and 0.74 for the one that uses word embeddings. The results obtained in the third experiment are revised only by one annotator. We finally compute from the annotations the percentage of incorrect and good translations, and the improvement over the baseline.

Table 5.2 shows the results of the manual evaluation. We observe that the combination of lexical chain detection using word embeddings with our Lexical Chain Translation Model (1) performs best. In particular, 81%-88% of the changes are correct translations

and, among them, 36%-48% are improvements over the baseline. Only 12%-19% of the changes are incorrect. Using GermaNet to detect lexical chains, the correctness decreases between 10% and 26%. Word embeddings may work better than lexical resources as they capture contextual information from the text, without relying on whether is defined in a resource. In those cases, where the resource does not provide a relation for two given words such as in idiomatic or metaphoric uses, the lexical chain cannot benefit from them.

The parameters length, density, and repetition have an impact on translation when using them to compute the strength of each lexical chain in the overall LCTM score. We see that the correctness of the translation output decreases approximately by 20% in all test sets when using the LCTM_{base} (i.e. the model that gives the highest weight to all lexical chains, ignoring the mentioned parameters) instead of the LCTM. Furthermore, the percentage of the improvements over the baseline decrease by half.

Some translation examples using our method are illustrated in figure 5.2. In the first example, the ambiguous German noun *Politik* gets correctly translated into *politics*. *Politik* is connected to *politischer* ("political") in the lexical chain, and therefore, *politics* is semantically more related to *political* than *policy*. Our method is also good at enforcing the translation of all words in the lexical chain, since an untranslated word will decrease the score of the translated lexical chain and, accordingly, the overall Lexical Chain Translation Model (LCTM) score (see example 2). In the last example, the method produces a wrong translation of the German word *Lohn* ("wage", "salary"), whereas the baseline translates it correctly. The word *Lohn* is linked to *erhöht* ("increase") and *Lohnerhöhungen* ("wage increases") in the lexical chain. Both words provide good context for the translation. However, our method incorrectly translates it into *pay*, whereas the baseline translates it correctly into *wage*.

In the third example, we observe that the method produces a different but equally good translation compared to the baseline. In the lexical chain, the German word *Rakete* is linked to another occurrence of the same word that is translated into *missile*. Since the highest similarity score is obtained when both translations are the same, our method encourages consistency (Carpuat, 2009, Carpuat and Simard, 2012), translating both into *missile*. Consistency is possible since we assume that there is only a unique sense per word in each document (Gale et al., 1992) as discussed in chapter 4. For example, both senses of the German *Decke*, *blanket* and *ceiling* in English, do not appear in the same text.

Figure 5.3 shows two more examples that illustrate the benefits and issues of consistent translation. These are special cases, where the word in the lexical chain is linked only to other occurrences of the same word. In the first example, we see that the word *Wahl*

Chain	***Politik*** → *politischer*
Input	Ich bin ein Neuling in der Prager **Politik**
Ref.	I'm a novice in Prague ***politics***
Baseline	I am a newcomer in the Prague ***policy***
Word Emb. & LCTM	i am a newcomer in the prague ***politics***
Chain	*erklärt* → ***meint*** → *meint*
Input	"Hier geht niemand vor Gericht", **meint** ...
Ref.	"Nobody will sue them here", ***said***...
Baseline	"Here is no one in court", ...
Word Emb. & LCTM	"Here is no one in court", ***says***...
Chain	*Rakete* → ***Rakete*** → *Motor*
Input	...technische Schäden an der **Rakete**
Ref.	...technical damage to the ***missile***
Baseline	...technical damage to the ***rocket***
Word Emb. & LCTM	...technical damage to the ***missile***
Chain	*erhöht* → ***Lohn*** → *Lohnerhöhungen*
Input	das sind ca. 1,1% mehr als sie jetzt für **Lohn** spenden.
Ref.	this is about 1.1% more than it spends on ***salaries*** right now.
Baseline	this is about 1.1% more than they now for ***wage*** donations.
Word Emb. & LCTM	this is about 1.1% more than they now for ***pay*** donations.

FIGURE 5.2: In these examples, the method produces a correct translation of the ambiguous word *Politik*, forces the translation of the German verb *meint*, and generates another good translation of *Rakete*. In the last example, the presented method incorrectly translates *Lohn* into *pay*, despite the context given by the lexical chain: *ehöht* ("increase") and *Lohnerhöhungen* ("wage increases").

is translated by the baseline into the wrong sense *choice*. Here, *Wahl* is linked to two other occurrences of the same word in the lexical chain, which are translated into the other sense *election*. Since the method obtains the highest score when the translations are the same, it either enforces all three occurrences to be translated into *election*, or *choice*. The LCTM score competes with the other models (e.g. language and translation model). The overall score when using the translation *choice* is then lower than when using *election* due to the other models, since *choice* does not fit in the local context of the other sentence.

In the second example, however, the method translates the wrong sense of *Verhältnis*. That is because the two senses of the word *Verhältnis*, *ratio* and *relationship*, appear in the same document. This fact violates the one-sense-per-discoures hypothesis (Gale et al., 1992), and when the only context provided by the lexical chain is the word itself, the method cannot disambiguate the senses.

Input	Er entschloss sich erst auf den letzten Moment, sich an der **Wahl** vor der letzten Hauptversammlung zu beteiligen
Ref.	He decided to participate in the ***elections*** before the last general meeting at the last moment
Baseline	He decided at the last moment, the ***choice*** of the last Hauptversammlung to participate
LCTM	He decided at the last moment, the ***election*** of the last Hauptversammlung to participate
Linked to	
Input	Pelta hatte schon vor der nicht realisierten Wahl ... versprochen, er wolle sich im Falle seiner **Wahl** auf die Nationalmannschaft ...
Ref.	Prior to the failed ***election*** ..., Pelta promised that if he was to win, he would focus on both the representation ...
LCTM	Pelta was not carried out before the election ... promised that he wanted to in the event of his ***election*** to the national team ...
Input	Er ist überzeugt, dass für die heutigen Probleme mit der **Wahl** die Euphorie verantwortlich ist, die zu Zeigen von Hašek herrschte.
Ref	He believes current problems with ***elections*** are caused by the euphoria there was in the time of Hašek's reign.
LCTM	He is convinced that, for today's problems, with the ***election*** of the euphoria is responsible, the times of Hašek there.
Input	Das **Verhältnis** der Länge der beiden erwähnten Finger ...
Ref.	The ***ratio*** of the length of those two fingers ...
Baseline	The ***ratio*** of the length of the two ...
LCTM	The ***relationship*** between the length of the two ...
Linked to	
Input	... dennoch halte er das **Verhältnis** zwischen der Fingerlänge und dem Krebsrisiko bisher nicht für "völlig erwiesen"
Ref.	... but in his opinion the ***relationship*** between the length of the fingers and the cancer does not seem to be "fully proven"
LCTM	... but it the ***relationship*** between the Fingerlänge and the risk of cancer has so far not been for "completely proven to be"

FIGURE 5.3: These examples show how the presented method behaves when a word in the lexical chain is linked to the same word in the text. In the first example, the German word *Wahl* is linked to other two occurrences of *Wahl* in the text. They both are correctly translated into *election*, and therefore, the LCTM gets a higher score when the first sentence is translated into the same term. This produces an improvement over the baseline that wrongly translates it into *choice*. In the second example, we observe that both senses of the word *Verhältnis* occur in the same document, forcing the first occurrence to be wrongly translated into *relationship*.

5.5 Length, Density and Repetition in Lexical Chains

The method presented in this chapter takes into account the relevance (or strength) of every lexical chain in the computation of the Lexical Chain Translation Model score.

The strength of a lexical chain is defined by its length (λ), repetitions among the words in the chain (β), and the ratio of the number of semantic relations to the total number of words in the text segment that it covers (i.e. density)(ρ). In this section, we tackle a study on the impact of each parameter (i.e. length, density, and repetition) on translation.

In section 5.4 we presented a version of the LCTM that gives the same importance to all lexical chains by setting all three parameters to their maximum value (i.e. 1.0). We observe that this approach underperforms compared to the original method, which shows the importance of making a distinction among the different lexical chains in a document.

The lexical chain detection method detects long, dense lexical chains with high repetition, which capture the meaning of a considerable portion of text, but also other less relevant lexical chains. The former needs to be translated correctly to preserve the meaning in the target. Weak lexical chains (e.g. those that are constituted by two or three distant words in the text) have less context in the chain itself, which may lead to translation errors. However, they are still valuable information. Note that the method can only improve the translation of the words covered by the lexical chains. Therefore, it is desirable to consider the highest number of semantically-similar words in the text as possible. Giving a different strength score to each lexical chain, ensures that the document translation does not get affected by a wrong translation candidate that belongs to a weak lexical chain.

To carry out the study on the three parameters that contribute to the computation of the strength, the implemented Lexical Chain Translation Model (LCTM) allows to define as input a weight for each of them in the range 0.0 to 1.0. The LCTM computes now the weight w of a lexical chain ch_i as the average of $w_\rho \cdot \rho_{ch_i}$, $w_\lambda \cdot \lambda_{ch_i}$ and $w_\beta \cdot \beta_{ch_i}$.

This approach allows us to test the impact on translation of each of them individually, or in other combination of weights. For example, if the weights of density and repetition are set to 0.0 and length to 1.0, the LCTM only considers the length to compute the strength of each lexical chain in a document.

The data used for this experiment is the same as detailed in section 5.3. The development test set newstes2010 is used to obtain the weight configurations that gives the best performance. These are then tested on newstest2011, newstest2012, and newstest2013.

TABLE 5.3: This table shows the results of performing grid search over the parameter values, using different weight configurations of length (λ), density (ρ), and repetition (β). The configuration that sets all weights to 0.0 would correspond to a baseline that ignores discourse knowledge. In contrast, all weights set to 1.0 is the configuration used in the original method. When only density is activated (c_2), the method gets the lowest value of translation changes in the output. The best configurations are c_{10} and c_{12} according to the manual evaluation and c_{13} according to the BLEU scores.

	λ	ρ	β	BLEU	+	++	−	mod
c_1	**0.0**	**0.0**	**1.0**	14.64	0.65	0.35	0.35	40
c_2	**0.0**	**1.0**	**0.0**	14.75	0.60	0.60	0.40	5
c_3	0.0	1.0	1.0	14.78	0.61	0.32	0.39	28
c_4	0.0	1.0	0.5	14.74	0.57	0.30	0.43	23
c_5	0.5	0.0	1.0	14.70	0.55	0.32	0.42	38
c_6	0.5	1.0	0.0	14.78	0.57	0.21	0.43	14
c_7	0.5	1.0	0.5	14.73	0.60	0.28	0.40	25
c_8	**1.0**	**0.0**	**0.0**	14.69	0.63	0.26	0.37	35
c_9	1.0	0.0	1.0	14.75	0.68	0.35	0.33	40
c_{10}	*1.0*	*0.0*	*0.5*	14.76	0.70	0.35	0.30	40
c_{11}	1.0	1.0	0.0	14.79	0.68	0.26	0.32	19
c_{12}	*1.0*	*1.0*	*0.5*	14.72	0.70	0.35	0.30	23
c_{13}	*1.0*	*0.5*	*0.5*	14.80	0.62	0.27	0.38	26
LCTM	1.0	1.0	1.0	14.74	0.68	0.36	0.32	28

Table 5.3 summarises the results obtained on the development set for all configurations of weights 0.0, 0.5, 1.0. Some configurations such as $\lambda = 0.5$, $\rho = 0.0$, and $\beta = 0.0$ and $\lambda = 0.5$, $\rho = 0.0$, and $\beta = 0.5$ do not appear in the table, since they are the same as $\lambda = 1.0$, $\rho = 0.0$, and $\beta = 0.0$ (c_8) and $\lambda = 1.0$, $\rho = 0.0$, and $\beta = 1.0$ (c_9), respectively.

The performance of each combination is measured by a manual evaluation and in BLEU scores. The manual evaluation is carried out as described in section 5.4. Table 5.3 also shows the number of changes on the translation due to each weight configuration. Note that the configuration $\lambda = 1.0$, $\rho = 1.0$, and $\beta = 1.0$ corresponds to the results obtained with the original method. Similarly, $\lambda = 0.0$, $\rho = 0.0$, and $\beta = 0.0$ corresponds to the baseline that ignores lexical chains.

According to the results in the table, density is the factor that has the lowest impact (in terms of translation changes) on the output. Indeed, c_2 causes five changes in contrast to 40 and 35 by c_1 and c_8, which account for only repetition and length, respectively.

The manual evaluation finds that the best translation output is achieved with the configurations c_{10} and c_{12}. These get slightly better results than the original method.

As we introduced in section 5.4, BLEU scores are not informative in these experiments as the fluctuations are very small ($\pm$ 0.1). However, we also consider in this experiment the configuration c_{13}, which gives the highest BLEU score, for testing.

TABLE 5.4: Manual evaluation results of the three different weight combinations of the parameters length, density and repetition compared to the original method (1) and the approach that ignores the parameters (2). The analysis shows the percentage of correct (+) and wrong translations (-), and the improvement over the baseline (++). There are no remarkable differences in the manual evaluation between them and the original method (1) performs best.

	newstest2011			**newstest2012**			**newstest2013**		
	+	−	++	+	−	++	+	−	++
Word Emb. & LCTM (1)	0.81	0.19	0.48	0.88	0.12	0.36	0.83	0.17	0.39
Word Emb. & LCTM$_{base}$ (2)	0.64	0.36	0.22	0.67	0.33	0.18	0.61	0.39	0.16
1 0 5 (3)	0.72	0.28	0.34	0.80	0.20	0.32	0.78	0.22	0.38
1 5 5 (4)	0.78	0.22	0.43	0.76	0.24	0.36	0.78	0.22	0.30
1 1 5 (5)	0.76	0.24	0.43	0.73	0.27	0.36	0.80	0.20	0.33

The configurations c_{10}, c_{12}, and c_{13} are then applied on newstest2011, newstest2012, and newstest2013. Table 5.4 shows the manual evaluation of the results obtained. We do not observe any remarkable difference between the three new configurations and the original method still produces the best translation quality.

As discussed earlier, density alone has a low impact on the output. Experimenting with other weight configurations does not lead to better results. Therefore, we conclude from this study that the three factors length, density, and repetition are equally relevant as the original method performs best.

5.6 Summary

In this chapter, we presented a method that utilises lexical chains to improve the quality of document-level SMT output, showing that the translation output improves when discourse knowledge is considered. Specifically, the method improves the translation of the words in the chains, keeping the semantic similarity from the source to the translation. Each lexical chain captures a portion of the cohesive structure of a document. It is therefore essential to ensure that the words in the lexical chains are well translated.

The method is divided into two steps that consist of detecting the lexical chains in the source and preserving the semantic similarity among the words in their counterpart target lexical chains. We therefore implemented an automatic detection of the lexical chains based on a manual approach proposed by Morris and Hirst (1991) and a feature function in the document-level decoder Docent (i.e the Lexical Chain Translation Model (LCTM)) that preserves the semantic similarity in the translated chains.

The novelty of this approach is that we use word embeddings instead of external lexical resources to deal with word similarity. In dictionary-based approaches, the resources are

used to detect the relations between the words in the lexical chains, but it is not relevant for this task to know the kind of relation that connects two words. In some cases (e.g. idiomatic or metaphoric uses), the dictionary does not provide a relation between two words if there is no relation assigned in its lexicon. Accordingly, our method computes the similarity between words using word embeddings regardless of the pre-defined set of senses from dictionaries.

The problem of word embeddings is that polysemic words are represented with a single word embedding that captures the information of all its meanings. Therefore, it is necessary to disambiguate words in the training phase to be able to detect the similarity between polysemic words. For this reason, we applied the approach described by Thater et al. (2011), which relies on syntactic information to differentiate a word that appears in different contexts.

Thater et al. (2011)'s approach relies only on sentence-level context, and so, it does not suffice if it needs context from previous sentences to correctly disambiguate. A way to overcome this issue and improve the performance of the lexical chain detection consists on combining word embeddings and a external lexical resource (e.g. GermaNet as in our experiments). This way, we would first try to identify pairs of words as semantically related using the lexical resource. That is, words that have a synonym, hypernym, hyponym, or sibling relation between them. If they are related, we would include them directly in the lexical chain. Otherwise, despite not having a relation defined in the dictionary, they could still appear in similar contexts, which also helps our model to improve the lexical choice. The word embeddings would be then used to detect those words that occur in similar documents by computing their similarity as we did in this chapter. Our experiments did not combine word embeddings and lexical resources, but we suggest to continue the research on lexical chains in this direction.

We assessed the performance of the lexical chain detection on the translation task. The manual evaluation of the results show that the proposed method improves between 36% and 48% of the changes over a baseline that does consider lexical chains or any document-level knowledge. The results of the method are also evaluated against the method proposed by Galley and McKeown (2003), which uses a dictionary instead of word embeddings.

The results of all experiments are manually evaluated, since fluctuations in BLEU scores are very small ($\pm$ 0.1). The main reason is that the method does not tackle the translation of all words in a document, but only the ones covered by lexical chains. From those words, only the ones that are ambiguous can improve over the baseline. Furthermore, often the translation proposed by the method was correct, but it did not match the reference. In those cases, the BLEU scores cannot reflect the improvement.

The method showed a bias for consistently translating the words in the chain. Since we assume the one-sense-per-discourse hypothesis (Gale et al., 1992), this it the preferable behaviour as discussed in chapter 4. Here, the method has the advantage that the Lexical Chain Translation Model competes with other feature functions during decoding. This way, when multiple occurrences of the same word are linked in a lexical chain, the decoder favours the consistent use of the translation that fits in all their contexts, avoiding translating them in the wrong sense.

When the one-sense-per-discourse hypothesis does not hold, different senses of the same word may end linked in the same lexical chain. This is a problem when each sense has a different translation in the target language. The method cannot then distinguish between different senses, translating incorrectly both into the same translation sense.

The lexical chains detected in the source differ from each other in length, density, and total of repetitions. To ensure that they have a different degree of impact on translation depending on their strength in the document, the LCTM takes that into account in the computation of the model score. Accordingly, the LCTM implements the computation of their strength based on the aforementioned three parameters. Morris and Hirst (1991) are the first to introduce the definition of lexical chain strength, although they did not use it in their experiments.

To assess the importance of distinguishing between lexical chains, we implemented a simpler version of the Lexical Chain Translation Model (i.e. LCTM$_{base}$) that gives the same strength value to all chains in the document. The experimental results showed that the method that uses the LCTM$_{base}$ performs worse that the original method in all test sets.

We then extended the study on strength of lexical chains and assessed different weight configurations of the parameters length, density, and repetition. Specifically, we defined the weight of each parameter (0.0, 0.5, or 1.0) as input of the LCTM. This allowed us to evaluate the impact on translation of each parameter independently (i.e. the weight of the parameter we want to evaluate is set to 1.0 and the others to 0.0) and other combinations (e.g. length 1.0, density 0.5, and repetitions 0.0).

The results showed that density is the parameter that has the lowest impact on translation: only 5 translation changes versus 35 and 40 for length and repetition, respectively. Furthermore, no other combination beat the results of the original method. Therefore, we conclude that all parameters are equally important in the computation of the strength of lexical chains.

Chapter 6

Document-level Neural Machine Translation

Neural MT has recently emerged as the new machine translation paradigm, reporting a translation performance competitive to state-of-the-art Statistical MT systems. However, there has been little attention on taking into account wider context than the sentence itself during the training and translation phases.

In this chapter, we address how to integrate discourse-knowledge in Neural Machine Translation and assess whether it improves the lexical choice of the translation systems. To do so, we continue with the research on the lexical chains approach described in chapter 5 and integrate them in the Neural Machine Translation system as additional input factors (section 6.1). Additionally, we evaluate how well Neural Machine Translation learns from sense labels and compare this sentence-level approach with our lexical chains model.

For the experiments, we focus on the German→French and German→English translation and evaluate our discourse-oriented method on the Word Sense Disambiguation task ContraWSD (Rios et al., 2017),[1] which is specially designed to test the lexical choice performance of Neural Machine Translation models on ambiguous German words (section 6.2). The experimental results show that while the baseline is good at predicting frequent word senses, our approach slightly improves the prediction of rare senses (section 6.3).

[1] https://github.com/a-rios/ContraWSD

6.1 Lexical Chains in Neural Machine Translation

In this section, we describe our method to provide the Neural MT encoder with discourse knowledge. Similarly to the method presented in chapter 5, we exploit lexical chains in the source as a means to make document-level semantic information available to the MT system. The method is divided into two steps: (1) to detect the lexical chains in the source and (2) to build the factors that will feed the NMT system with the information from the detected lexical chains.

6.1.1 Computation of Semantic Similarity

Our method exploits lexical chains from the source document, and in order to detect them, we need to compute the semantic similarity between pairs of words in a document as in section 5.2. Unlike the experiments presented in chapter 5, where we use the lexical chains only for testing, we use them here also for training the Neural Machine Translation system, which considerably increases the amount of data to be processed. We therefore were not able to use the method presented in section 5.2.3, which distinguishes between word senses through syntactic information as it was computationally very expensive.

In this chapter, we use SenseGram (Pelevina et al., 2016)[2] to deal with sense embeddings, which allows us to learn sense embeddings and apply them to disambiguate the words in our data. Their experimental results show that the method achieves a F-score of 0.840 on the sense-balanced TWSI dataset (Biemann, 2012).[3] Additionally, the performance of this approach is comparable to state-of-the-art unsupervised WSD systems in the SemEval-2013 task 13 (Jurgens and Klapaftis, 2013).

The method to learn the sense embeddings with SenseGram consists of four steps outlined as follows: it first learns word embeddings using the word2vec toolkit (Mikolov et al., 2013). It then uses the word embeddings to build a word similarity graph, where each word is linked to the 200 nearest neighbours (i.e. words with the highest cosine similarity). Next, it induces a sense inventory, where each sense is represented by a cluster of words. For example, the sense of *table-furniture* is represented with the word cluster *desk*, *bench*, *dining table*, *surface*, and *board*. The sense inventory of each word is obtained through clustering the ego-networks of its related words. Finally, the method computes the sense

[2] https://github.com/tudarmstadt-lt/sensegram
[3] https://www.lt.informatik.tu-darmstadt.de/de/data/twsi-turk-bootstrap-word-sense-inventory/

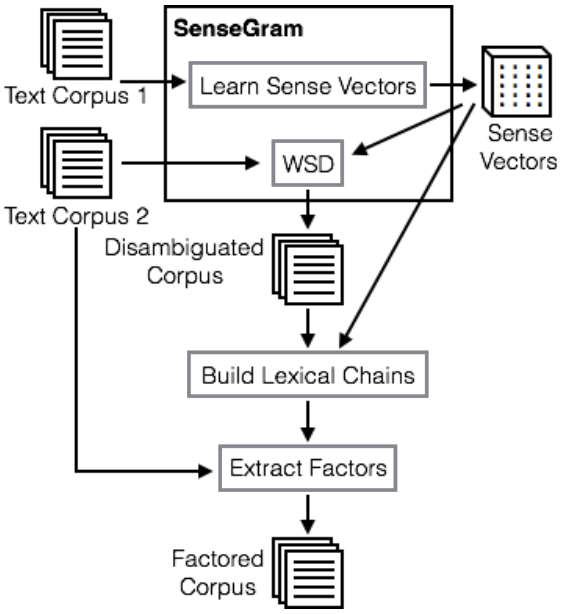


FIGURE 6.1: Diagram of the pipeline used to compute the factored corpus, where words that are detected in a lexical chain contain the linked words in the chain as input features. The factored data is then used to train and test the NMT system.

embedding of each word sense by averaging the vectors of the words in the corresponding cluster.

Once the sense embeddings are learned, we use them to disambiguate the words in our data. SenseGram allows us to disambiguate a word given the sentence in which it appears as context. We then label all content words in the data with their corresponding proposed sense. The following German sentence results from applying WSD with SenseGram on every content word: *Ich freue#1 mich, dass meine Vorschläge#0 eine so positive#3 Resonanz#0 gefunden#5 haben*, where the number represents the identifier of the word sense learned with SenseGram. The disambiguated words are finally used to detect the lexical chains.

To detect the lexical chains, we need to compute the semantic similarity between word senses. To do so, we calculate the cosine distance between their sense embeddings as detailed in section 5.2.3.

6.1.2 Annotation of Lexical Chains as Input Features

To provide the NMT encoder with semantic information from lexical chains, we use the method presented in section 5.2.1 to detect lexical chains in the source document and the sense embeddings learned with SenseGram to compute the semantic similarity, as

detailed in section 6.1.1.[4] Next, we represent this discourse knowledge in the input as a combination of features (Alexandrescu and Kirchhoff, 2006, Sennrich and Haddow, 2016). Accordingly, each word in the lexical chain is represented together with its linked words (i.e. previous and next words in the chain) as factors. For example, if the German word *Absatz* ("heels", "paragraph", or "sales") is linked in the lexical chain to *Wirtschaft* ("economy") and *Verkauf* ("sale"), it is then represented as *Absatz*|*Wirtschaft*|*Verkauf*. The vector representation of *Absatz* becomes then the combination of each feature's embeddings (i.e. the embeddings of *Absatz*, *Wirtschaft*, and *Verkauf*). Figure 6.1 gives an overview of the stages in the process of obtaining the factored data that we use as input to train the NMT system and to translate the test sentences.

In our experiments, we encode all words from the input data via joint byte pair encoding (BPE) as indicated in section 1.3.4. For example, the German *unbedachten* ("thoughtless") results in our data as the two subword units: *unbe@@ dachten*.[5] The additional input features are then replicated in each of the subword units as follows. On the one hand, if the current word (e.g. *unbedachten*) does not belong to any lexical chain, the additional input features are the subword unit itself (e.g. *unbe@@*|*unbe@@*|*unbe@@ dachten*|*dachten*|*dachten*). On the other hand, if the current word belongs to a lexical chain, the input features are then the linked words (maximum two) in its lexical chain (e.g. *unbe@@*|w_x|w_y *dachten*|w_x|w_y, where w_x and w_y are two words linked to *unbedachten* in the lexical chain). If the word is linked to only one word in the lexical chain, we also add the corresponding subword unit (e.g. *unbe@@*|*unbe@@*|w_x *dachten*|*dachten*|w_x, where w_x is the linked word in the lexical chain)

Since the input features that result from a lexical chain (w_x and w_y) could have a different number of subword units than the input word, we do not apply BPE to them and add their complete word form as input feature. For example, if *unbedachten* is only linked to *Lächerlichkeit* ("ridiculousness") in the lexical chain, we then obtain *unbe@@*|*unbe@@*|*Lächerlichkeit dachten*|*dachten*|*Lächerlichkeit*.

6.2 Setup of the Word Sense Disambiguation Task

We build German→English and German→French NMT systems with document-level context from the lexical chains represented as input features in the data (section 6.1.2).

[4] The code to detect lexical chains is available at https://github.com/lmascarell/lexCH

[5] Note that BPE encoding appends @@ to all subword units except the last one to indicate the end of each sequence of subword units.

Source	Fest steht, daß das Haus den ersten **Absatz** angenommen hat.
Reference	Clearly the Assembly has adopted the first **paragraph**.
Contrastive	Clearly the Assembly has adopted the first **heel**.
Contrastive	Clearly the Assembly has adopted the first **sales**.
Source	Er hat zwar schnell den Finger am **Abzug**, aber er ist eben neu.
Reference	Il a la **gâchette** facile mais c'est parce qu'il débute.
Contrastive	Il a la **soustraction** facile mais c'est parce qu'il débute.
Contrastive	Il a la **déduction** facile mais c'est parce qu'il débute.
Contrastive	Il a la **sortie** facile mais c'est parce qu'il débute.
Contrastive	Il a la **rétraction** facile mais c'est parce qu'il débute.

FIGURE 6.2: Contrastive translations of a German sentence containing the ambiguous word *Absatz* and *Abzug* in English and French, respectively.

To assess how well senses perform compared to this lexical chains approach, we additionally train a system for each language direction, where each word in the data contains its sense obtained with SenseGram (see section 6.1.1) as an additional feature (e.g. *Vorschläge*|*Vorschläge#0*). This approach is not discourse oriented, although it takes into account a wider context than phrase-based systems, since SenseGram uses the whole sentence to disambiguate the sense of each word. In section 1.3.4, we detail the configuration of the NMT systems.

We then assess the performance of the systems on the Word Sense Disambiguation task ContraWS compared to a baseline NMT system that does not integrate any discourse or sense information. In this task, we use a test set of contrastive translations that is designed to test the performance of Neural MT models on lexical choice. The test set is a collection of German sentences that contain an ambiguous German word, its reference, and at least one automatically generated contrastive translation, where the translation of the ambiguous word is replaced with one of its other senses. Figure 6.2 shows an example of the ambiguous German word *Absatz*, which can be translated into English *heel*, *sales*, and *paragraph*, and *Abzug* ("deduction", "withdrawal", "discount") with its French translations.

To assess the capability of each system's model at distinguishing different word senses, we let the model score the reference and each of the contrastive translations. We then count it as a correct decision, if the reference score is higher than all the other contrastive translation scores.

6.2.1 Training and Test Corpora

The parallel training data that we use to build the German→English and German→French NMT systems comes from OPUS (Tiedemann, 2012b) and consists of about 2.1 million sentence pairs from Europarl (v7) and ∼207K sentence pairs from News Commentary (v11) for both language directions. As discussed in section 1.3.2, we considerably improve the performance of a NMT system by adding a larger amount of training data (e.g. Common Crawls). Indeed, the Edinburgh systems (Sennrich et al., 2016a), which achieved the best performance at the WMT'16 translation task (Bojar et al., 2016) in several language pairs, use ∼4M sentence pairs for training the winner German→English and English→German systems. However, we are restricted to corpora that contain document boundaries, as we need to detect the lexical chains in each document.

Our method uses sense embeddings to detect lexical chains in the source document. We therefore train a skip-gram 300-dimensional model in German using the SenseGram tool as described in section 6.1.1. Our data to learn the senses comes from SdeWaC (Faaß and Eckart, 2013) (∼768M words), Common Crawls (∼775M words), Europarl (∼47M words), and News Commentary (∼6M words).

The German-English test set of ContraWSD is built upon 83 different word senses, where each sense is represented by 100 instances or the total amount if they occur less than 100 times in the data. In total, the test set contains 7,243 sentence pairs with an average of 3.5 contrastive translations. Similarly, the German-French test set contains 6,746 sentence pairs and a total of 71 senses, with an average of 2.2 contrastive translations.

The sentence pairs are extracted from the Credit Suisse News Corpus (∼95.7K sentences),[6] the United Nations Parallel Corpus (∼166K sentences) (Ziemski et al., 2016), the WMT test and development sets from 2006 to 2016 (∼28K sentences),[7] and corpora from OPUS (Tiedemann, 2012b) such as Global Voices (∼57.4K sentences),[8] Books (∼51.4K sentences),[9] and the EU Bookshop Corpus (∼9.3M sentences).[10] To evaluate our models, we exclude newstest2010, since we use it as a development set to train the NMT systems.

We also evaluate the performance of our systems on automatic Bleu scores. For this, we use the test sets newstest2009, newstest2011, newstest2012, and newstest2013 for

[6] https://pub.cl.uzh.ch/projects/b4c/de/
[7] http://www.statmt.org/wmt17/translation-task.html
[8] http://opus.lingfil.uu.se/GlobalVoices.php
[9] http://opus.lingfil.uu.se/Books.php
[10] http://opus.lingfil.uu.se/EUbookshop.php

TABLE 6.1: Word sense disambiguation accuracy

German→English	Accuracy
NMT baseline	0.7095
NMT sense labels	0.7138
NMT lexical chains	0.7034
human	≈0.96
German→French	Accuracy
NMT baseline	0.7023
NMT sense labels	0.6998
NMT lexical chains	0.7083
human	≈0.93

testing, as they are available in both German→English and German→French language directions.

We evaluate the performance of the systems on a Word Sense Disambiguation task, using the ContraWSD test set, and compute their BLEU scores on the translation of the test sets newstest2009, newstest2011, newstest2012, and newstest2013. These systems are presented as baselines for future experiments, as the goal of the ContraWSD test set is to support future research on Word Sense Disambiguation in NMT. Since other groups may want to test single models on the test sets, and we expect high variability between checkpoints during the training of the NMT model, we assess the performance of each system using the model that gives the highest BLEU score during training.

We summarise the accuracy of the systems at scoring the reference higher than the contrastive translations in table 6.1. Both German→English and German→French baselines achieve a 0.70 accuracy on the WSD task. We notice a small improvement using the system trained with sense labels for German→English, and with lexical chains for German→French.

The table also shows the accuracy of a small-scale human evaluation (one annotator per language pair) on a random sample of the test set with 100 to 150 sentence pairs. The annotation is provided with the test set and it is performed at sentence level, without discourse context. The results on the annotation task show that some ambiguities are even difficult to resolve for humans without document context, as shown in example 6.1.

(6.1) **Source:** Sehen Sie die *Muster*?
Reference: Do you see the *patterns*?
Contrastive: Do you see the *examples*?

TABLE 6.2: Accuracy of the word sense prediction by frequency of the senses in the training set. The *#Senses* column represents the number of word senses found in that frequency range. The table shows that the baseline is reliable for frequent word senses, but not for the rare ones, where the proposed systems trained on lexical chains and sense labels help to slightly improve the accuracy.

	German→English				German→French			
Train Freq.	#Senses	baseline	sense labels	lexical chains	#Senses	baseline	sense labels	lexical chains
>10000	2	0.9840	0.9840	0.9840	2	0.9900	0.9900	0.9900
>5000	7	**0.9639**	0.9534	0.9459	1	1.0000	1.0000	0.9900
>2000	4	**0.9386**	0.9284	0.9284	3	0.7375	**0.7725**	0.7150
>1000	6	0.8598	**0.8632**	0.8427	3	0.9333	**0.9367**	0.9167
>500	8	**0.7410**	0.7308	0.7090	6	0.8260	0.8260	**0.8361**
>200	17	0.7800	0.7734	**0.7900**	16	0.8444	**0.8475**	0.8406
>100	9	0.6058	0.6095	**0.6156**	9	**0.7544**	0.7456	0.6933
>50	8	**0.7899**	0.7645	0.7630	6	0.5160	0.5200	**0.6420**
>20	9	0.4055	**0.4521**	0.3945	8	0.5276	0.5430	**0.5469**
0-20	14	0.3127	**0.3664**	0.3237	17	0.4924	0.4611	**0.5156**

TABLE 6.3: Translation examples of the German→English system trained on lexical chains, where the system improves the lexical choice of the ambiguous German words.

Source	Wenn demokratische Wähler der Vorwahlen gefragt werden, wer für die Partei ihre zweite **Wahl** wäre...
Baseline	If democratic voters are asked for pre-election, who would be the second **election** for the party...
Lexical Chains	When democratic voters are asked for the primaries, who for the Party would be their second **choice**...
Reference	When Democratic primary voters are asked who would be their second **choice** for the party's nomination...
Source	Wenn ein Spieler mit dieser Art von Verletzung den **Platz** verlässt, sind alle sehr traurig darüber.
Baseline	When a player leaves **room** for injury, everyone is very sad about it.
Lexical Chains	When a player leaves the **place** with this kind of violation, everyone is very sad.
Reference	If any player goes off with this kind of injury, everybody is very sad about it.
Source:	Gesundheitliche Schäden bis hin zu Essstörungen, eine weitere weitverbreitete **Annahme**, habe er nicht erlitten.
Baseline:	Health damage to eating disorders, which is more widespread, has not suffered.
Lexical Chains	Health damage to eating disorders, another widespread **assumption**, did not happen.
Reference	He has not suffered any damage to his health or eating disorders - another wide-spread **assumption**.
Source	Ein erneut enttäuschender PC-**Absatz** sowie eine schwächere Nachfrage...
Baseline	Another disappointing PC **paragraph**, as well as a weaker demand...
Lexical Chains	Once again, a more disappointing PC **sales** and the weaker demand...
Reference	Repeatedly disappointing **sales** of PCs as well as a slow demand...
Source	**Wahl**-2016: Hillary Clintons Vorsprung vor Bernie Sanders in nationaler Umfrage halbiert.
Baseline	**Post**-2016: Hillary Clinton's margin from Bernie Sanders-ten years of national survey.
Lexical Chains	**Election** 2016 - Hillary Clinton's withdrawal from Bernie Sanders in a national poll halved.
Reference	**Election** 2016: Hillary Clinton's lead over Bernie Sanders cut by half in national poll.
Source	Andererseits erlauben dieselben Staaten es Mitgliedern..., die ausgestellten **Karten** für Wahlen zu benutzen.
Baseline	On the other hand, the same states allow for members... to use the **maps** issued by these clubs.
Lexical Chains	On the other hand, the same states allow... to use the **cards** issued by these clubs for elections.
Reference	On the other hand, these same States allow... to use the **cards** issued by these clubs when they vote.

6.3 Evaluation of the Systems on the WSD task

In table 6.2 and figure 6.3, we show a more fine-grained evaluation of the results. Here, we group the senses by their frequency in the training data from 0-20 occurrences to more than 10,000 and present the accuracy scores for each group. For German→English, we observe that all models achieve an accuracy higher than 90% on words that occur more than 2,000 times in the training data. In contrast, for German→French, the accuracy of all systems on the frequency classes >5,000 and >10,000 is very close to 100%. For both languages pairs the accuracy of the baseline correlates with the frequency of the sense in the training data, and the more infrequent they are, the lower the accuracy. Thus, the baseline obtains only 31% accuracy for German→English and 49% for German→French on rare words such those seen 0-20 times during training, as indicated in table 6.2.

Since the baselines already make good predictions on frequent word senses, the lexical chains and sense systems do not show any improvement on those classes. However, they show a slight improvement over the baseline on rare and less frequent word senses. Specifically, the German→English system trained with the sense labels improves the accuracy by 0.43% over the baseline, and the German→French system with lexical chains beats the baseline by 0.6% points.

Despite the tendency of decreasing the accuracy of the systems as we move on to less frequent senses, we observe in figure 6.3 that the accuracy in some frequency groups are

FIGURE 6.3: Accuracy of the German→French (left) and German→English (right) word sense prediction grouped by frequency of the senses in the training set. The Figure shows that the baseline is not reliable for rare word senses, such as those occurring less than 20 times in the training data, where the proposed systems trained on lexical chains and sense labels help to improve the accuracy of the German→French and German→English, respectively.

TABLE 6.4: Average BLEU scores on newstest 2009-2013

	NMT baseline	NMT sense labels	NMT lexical chains
German→English	17.1	16.9	17.1
German→French	14.6	14.6	14.7

surprisingly high or low. That is the case of the >2,000 group for German→French, for example, where the accuracy is lower than in the less frequent group >1,000. This is due to the small number of word senses in those groups (i.e. only three in these two groups, as shown in table 6.2). Accordingly, if at least one of the senses in these groups gets systematically incorrectly or correctly predicted, the accuracy of the overall group considerably decreases or increases, respectively.

Table 6.4 shows the average BLEU scores of the systems on the test sets newstest2009, newstest2011, newstest2012, and newstest2013. The fluctuations in the scores are small, and only the German→English lexical chain system decreases by 0.2 BLEU points. In table 6.3, we list some examples, where the lexical chain model for German→English improves the lexical choice of the ambiguous word.

6.4 Summary

Neural networks have emerged as a revolutionary new paradigm in MT. Indeed, Neural MT systems achieved comparable and even better performance for some high-resource language pairs than the state-of-the-art SMT systems in the latest machine translation competitions (Luong and Manning, 2015, Sennrich et al., 2016a, Neubig, 2016).

MT research has been growing in this direction to solve the issues detected on NMT output, such as producing fluent translations that are unrelated to the source sentence or translating low-frequency verbs that are highly inflected (Koehn and Knowles, 2017). However, integrating discourse knowledge in NMT has received little attention so far.

In this chapter, we presented a method to integrate document-level context into NMT from lexical chains. We focused on the German→English and German→French language directions, and evaluated the performance of the systems on a test set of ambiguous German words, specially designed for NMT, which has a strong focus on difficult cases. To detect the lexical chains, we used the same method described in section 5.2.1, and we then included the detected lexical chains as additional factors in the input data. Additionally, we built two systems (one for each language direction) trained on the sense label of each word. Although this approach is not discourse-related, we wanted to

assess how well Neural MT learns from word senses and compare it to our lexical chains approach.

The experimental results showed that NMT resolves well the lexical choice of frequent words senses, even without any additional discourse or sense knowledge, but not infrequent word senses such as those seen less than 20 times in the training data. In those cases, we observed that the inclusion of sense labels slightly improved the lexical choice from German into English by 0.43% points, and the lexical chains increases the accuracy by 0.6% points for German→French.

The human evaluation revealed that document context is necessary to solve this task, and even though the reported improvements are small, we believe that future research needs to focus on how to integrate discourse context in NMT to improve lexical choice.

Chapter 7

Conclusions

Machine Translation systems translate sentence by sentence, ignoring context information from previous translated sentences or the whole document. This unawareness of the discourse leads to translation errors, since the context within a sentence is often not enough to disambiguate the translation of words, and the systems need information that cross sentence boundaries to make good lexical choices.

In this thesis, we aimed at improving the lexical choice of Machine Translation systems, integrating discourse knowledge. Specifically, we developed methods and performed experiments to integrate document-level context in Statistical and Neural MT decoders. In the following, we summarise the conclusions of this thesis, answering the four motivating questions introduced in section 1.1.

Research question 1: ***How important is the use of discourse knowledge to improve lexical choice compared to the local context provided by the surrounding words?***

The importance of discourse context in the translation process was the first thing to evaluate in this thesis. The intuitive answer is that discourse improves lexical choice, since human translators do not deal with each sentence in isolation but with the whole document to select appropriate word senses in the discourse.

To measure the importance of document context, we ran a discourse error analysis on the translation of different text genres, such as news articles, movie subtitles, transcribed talks, proceedings of the European Parliament and essays on the alpine domain, which we detailed in chapter 3. Specifically, we detected discourse-related errors and annotated whether they could be solved using local context or discourse context. For the experiments, we chose Germanic languages in the source (English and German) and

latin languages in the target (Spanish and French), and we trained the corresponding English→Spanish and German→French in-domain systems.

In our analysis, we found that nouns particularly benefit from discourse knowledge, across all genres, and adjectives are better translated using local context. The reason is that adjectives are usually positioned close to the noun they describe, which provides more local context. Of course, a bad translation of a noun in the document context is detrimental to the translation of the corresponding adjective.

In general, we found that discourse and local context are equally important to improve the lexical choice in both language directions. These findings suggest that state-of-the-art systems that operate at sentence-level also profit from integrating inter-sentential context information.

Research question 2: ***What kind of inter-sentential context information is useful to improve lexical choice in Machine Translation, and how can it be integrated into Machine Translation?***

In the same analysis mentioned in the previous research question, we also reflected on the way discourse reduces the translation errors in order to be able to model this knowledge into Machine Translation (chapter 3). We found that the vast majority of these errors can be solved either by encouraging consistent translations or producing a translation that is semantically similar to the words in the topic of the corresponding text fragment.

Translation consistency has been addressed in the literature as a way to improve lexical choice. Basically, if a word occurs several times in a document and at least one of them is well translated (e.g. due to better local context), the other mistranslations of the same word can be fixed by using the same correct translation throughout the document. Besides repetition, which is a way to achieve lexical cohesion in a document, we also recognised referential links between nouns, whose lexical choice is improved by enforcing translation consistency. Specifically, heads of nominal compounds that are later used to refer back to the compound itself can benefit from the translation of the compound, as the compound has less translation variants, and therefore, it is less ambiguous. In the analysis from chapter 3, we found only a couple of compound-reference examples. However, the references could not profit from their compound translation, because the compounds were not translated.

In chapter 4, we described several experiments to apply consistency for both repeated nouns and references to compounds. To consistently translate repeated nouns, we first needed to identify which of the repeated nouns is correctly translated. Thus, we trained classifiers to make such prediction based on syntactic and semantic features, where the

latter are extracted from local context and discourse context. In both cases, we experimented with several ways to produce consistent translations, such as by plugging the preferable translation into the decoder or by automatically post-editing the translation output. In *research question 3*, we continue with the discussion on consistency.

To disambiguate a word, it helped to look at the semantically-related words in the text fragment of the translation error. We therefore explored the use of chains of semantically-similar words (i.e. lexical chains) to improve translation and integrated them into a document-oriented decoder in chapter 5 and Neural Machine Translation in chapter 6. (We address discourse in NMT in *research question 5*). The discourse-aware model developed in chapter 5 encourages high semantic similarity between the translation of the words in the lexical chains.

In our experiments, we computed the quality of the translation output of our systems in terms of BLEU scores, but the improvements were always too small to get sufficient insight into the performance. BLEU is an automatic metric that is widely used in the Machine Translation community as a measure of a system's performance, but its n-gram matching approach to a single reference does not allow to capture the whole range of translation possibilities. Indeed, Martínez Garcia et al. (2017) state that "the usual automatic MT evaluation metrics are mostly insensitive to the changes introduced by our document-based MT system."

We then carried out manual evaluations that gave us a better understanding of the performance of our method. The improvement of our method over the baseline was then noticeable, but small, since the lexical choice errors that we tackled do not occur frequently, as analysed in chapter 3. The consistency experiments in chapter 4, for example, showed that the margin of improvement between the baseline and the oracle translations was very small. Nevertheless, we believe it is important to address these issues, as incorrect lexical choices lead to meaningless translations even if they are infrequent.

In addition, we noticed in the literature and also in our experiments on translation consistency that the biggest gains are reported for translations between English and Chinese. This fact shows that it is more difficult for the baseline to produce relatively good translations on a pair of typologically distant languages, and it is therefore easier for a system that integrates a discourse-aware model to outperform the baseline.

Research question 3: ***Is translation consistency desirable in the output of Statistical Machine Translation?***

Translation consistency has been the focus of several studies in the literature. Carpuat (2009) and Carpuat and Simard (2012) report that SMT systems translate consistently compared to human translations, and translation inconsistencies lead to incorrect lexical

choices more often than consistent translations do. However, it is difficult to determine whether consistency is desirable or not, since the enforcement of strict consistency may negatively affect fluency (Carpuat and Simard, 2012, Guillou, 2013). Guillou (2013) then suggests that consistency should be selectively enforced. For example, she found nouns to be a good target for consistent translation, across all genres.

Following Guillou (2013)'s findings, we experimented with consistent translation of only nouns in chapter 4. We did not actually tackle all nouns, but only those under specific conditions, such as repeated pairs of nouns and references to compounds. Our goal was to avoid too much consistency in the translation, so that it negatively affects the fluency, encouraging consistency only when it is expected. The results showed a small improvement in translation quality, and, even after narrowing the problem to these very specific scenarios, we still found that not all consistent translations were necessary, since the initial translations were already good translation candidates.

From the results of our experiments, we conclude that consistency should be only applied when there is an ambiguity issue, such as in the translation of polysemic words that have several translations in the target language. In other cases, lexical variability would be preferable. However, as Lyons (1968) states, words that can be interchanged, because they have the exact same meaning, are extremely rare. Human translators can better judge whether two translations of the same word fit in the context of the document. For Machine Translation this is a great challenge, and it is safer to use a correct translation repeatedly.

Research question 4: ***Can Neural Machine Translation profit from discourse context, and how can it be integrated?***

During the last year of the development of this thesis, Neural Machine Translation achieved a substantial improvement in translation quality over the state-of-the-art phrase based SMT systems for a number of high-resource language pairs. The NMT architecture still deals with sentences in isolation, and there was no research on integrating discourse context in NMT by the time we addressed this issue.

Our method to integrate discourse context into NMT was inspired by the work of Sennrich and Haddow (2016), which integrates additional linguistic information, such as morphology, part-of-speech tags, and syntactic dependency labels, as input features. Similarly, we obtained the lexical chains in the source using the method presented in chapter 5 and then added the connected words in the chain as additional features.

We assessed the performance of our German→French and German→English systems on a Word Sense Disambiguation task, which has a strong focus on difficult cases and is specially designed for the evaluation of NMT systems. The experimental results showed

that the NMT baseline performed well with the disambiguation of frequent word senses in the training data, but poorly for infrequent word senses that appear less than 20 times in the training data. In those infrequent cases, the German→French system that contains discourse knowledge from lexical chains outperformed the baseline.

An additional problem to consider for experiments on discourse in NMT is that we need a considerable amount of data that follows a document structure, which is not as widely available as other corpora consisting of random parallel sentences. In the experiments presented in chapter 6 we used a total amount of 2M parallel sentences mostly from Europarl. For Statistical Machine Translation, a training corpus consisting of 2M sentences per language is large enough to achieve a good state-of-the-art system. However, the best NMT systems reported in the WMT'16 translation task (Bojar et al., 2016) used around 4M parallel sentences. Indeed, Koehn and Knowles (2017) reported that the quality of NMT systems depends on the amount of training data even more than Statistical MT, and so, the more data we use to train the system, the better performance it can achieve.

Representing the discourse knowledge as additional input features is not the only way to integrate discourse into Neural Machine Translation. Indeed, later this year, Wang et al. (2017) and Jean et al. (2017) proposed an extension of the state-of-the-art attention-based NMT architecture (Bahdanau et al., 2015). In particular, Wang et al. (2017) modifies the architecture by adding a hierarchy of Recurrent Neural Networks to summarise the discourse context, reporting a considerable improvement in the translation quality from Chinese into English. Jean et al. (2017) evaluated the performance on a pronoun prediction task, which included the German-English and French-English language paris, and got mixed results. As discussed in *research question 2*, the results suggest that it is easier to achieve greater improvements for languages pairs that are are very distant and linguistically dissimilar, such as Chinese and English.

7.1 Future Research

There are three major lines of research in the future work of this thesis: the enforcement of consistent translation only when the inconsistent translation leads to wrong lexical choice, the improvement of lexical chain detection and its evaluation, and a deeper investigation of discourse in Neural Machine Translation. In the following, we discuss each of these aspects in detail.

Translation consistency has been deeply studied over the last years, resulting in mixed results on whether it should be enforced. However, these studies mostly tackle consistency in general, jeopardising the fluency of the translation output. We concluded from our experiments that consistency should be only applied when there is an ambiguity issue. That is, when a polysemic word has different translations in the target language and the decoder is using the translation in the wrong sense. For example, the German *Absatz* in the sense of sales should not be translated into the other senses "heel" or "paragraph", but "turnover" should be accepted, as it refers to the same sense. Future research on translation consistency should focus on accurately predicting when to apply a consistent translation to improve the quality of the translation output, leaving unchanged non-conflicting translations that are in the right sense.

In chapter 5 and chapter 6, we integrated discourse knowledge into Machine Translation systems using lexical chains, which are chains of semantically-similar words in a given document. The evaluation of lexical chains is difficult, as there is not just a single valid constellation of lexical chains in a document, and so we usually evaluate them on a specific task, such as word sense disambiguation or the translation task. In our experiments, the quality of the lexical chains is key to the improvement of lexical choice in translation: words that are not detected as part of a lexical chain (and they should) cannot be improved by our models or cannot help to improve the translation of other words in the chain. Thus, it is important to investigate how to improve lexical chain detection and also analyse the performance of external lexical resources in combination with our detection method, which uses word embeddings.

The last line of research in our future work is the integration of discourse knowledge into the architecture of Neural Machine Translation systems and the investigation of their benefits. Neural MT shows potential for dealing with discourse compared to SMT, as the attention mechanism in Bahdanau et al. (2015)'s architecture is already able to properly handle longer-distance dependencies within the sentence. We presented a method to integrate document-level information from lexical chains as additional input features in chapter 6. It remains for future experiments to develop a method that exploits the attention mechanism to integrate the lexical chains into the NMT architecture. In conclusion, from our experiments and the results reported by Wang et al. (2017) and Jean et al. (2017), we believe that NMT can benefit from discourse knowledge and that the NMT community should bear it in mind for future studies.

Bibliography

Alexandrescu, A. and Kirchhoff, K. (2006). Factored Neural Language Models. In *Proceedings of the Human Language Technology Conference of the NAACL*, pages 1–4, New York, NY, USA.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3^{rd} International Conference on Learning Representations*, San Diego, CA, USA.

Barzilay, R. and Elhadad, M. (1997). Using Lexical Chains for Text Summarization. In *Proceedings of the Workshop on Intelligent Scalable Text Summarization*, pages 10–17, Madrid, Spain.

Bentivogli, L., Bisazza, A., Cettolo, M., and Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. In *Proceedings of the 15^{th} Conference on Empirical Methods in Natural Language Processing*, pages 257–267, Austin, TX, USA.

Biemann, C. (2012). Turk Bootstrap Word Sense Inventory 2.0: A Large-Scale Resource for Lexical Substitution. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, pages 4038–4042, Istanbul, Turkey.

Bojar, O., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno Yepes, A., Koehn, P., Logacheva, V., Monz, C., Negri, M., Neveol, A., Neves, M., Popel, M., Post, M., Rubino, R., Scarton, C., Specia, L., Turchi, M., Verspoor, K., and Zampieri, M. (2016). Findings of the 2016 Conference on Machine Translation. In *Proceedings of the 1^{st} Conference on Machine Translation*, pages 131–198, Berlin, Germany.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1):5–32.

Bubenhofer, N., Volk, M., Klaper, D., Weibel, M., and Wüest, D. (2013). Text+Berg-Korpus (Release 147_v03). Digitale Edition des Jahrbuch des SAC 1864-1923, Echo des Alpes 1872-1924 und Die Alpen 1925-2011.

Carpuat, M. (2009). One Translation per Discourse. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 19–27, Boulder, CO, USA.

Carpuat, M. and Simard, M. (2012). The Trouble with SMT Consistency. In *Proceedings of the 7th Workshop on Statistical Machine Translation*, pages 442–449, Montréal, Canada.

Cartoni, B., Gesmundo, A., Henderson, J., Grisot, C., Merlo, P., Meyer, T., Moeschler, J., Zufferey, S., and Popescu-Belis, A. (2011a). Improving MT Coherence through Text-Level Processing of Input Texts: the COMTIS Project. In *Proceedings of the Tralogy - Translation and Natural Language Processing / Traduction et traitement automatique des langues*, Paris, France.

Cartoni, B., Zufferey, S., Meyer, T., and Popescu-Belis, A. (2011b). How Comparable are Parallel Corpora? Measuring the Distribution of General Vocabulary and Connectives. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora*, pages 78–86, Portland, OR, USA.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the 16th Conference of the European Association for Machine Translation*, pages 261–268, Trento, Italy.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., and Federico, M. (2015). The IWSLT 2015 Evaluation Campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Chen, X., Liu, Z., and Sun, M. (2014). A Unified Model for Word Sense Representation and Disambiguation. In *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*, pages 1025–1035, Doha, Qatar.

Chiang, D. (2005). A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 263–270, Stroudsburg, PA, USA.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, Doha, Qatar.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scale. *Educational and Psychological Measurement*, 20:37–46.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20(3):273–297.

Faaß, G. and Eckart, K. (2013). SdeWaC – A Corpus of Parsable Sentences from the Web. In *Language Processing and Knowledge in the Web*, volume 8105, pages 61–68. Springer Berlin Heidelberg.

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-1955. Blackwell.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One Sense per Discourse. In *Proceedings of the Workshop on Speech and Natural Language*, pages 233–237, Harriman, NY, USA.

Galley, M. and McKeown, K. (2003). Improving Word Sense Disambiguation in Lexical Chaining. In *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pages 1486–1488, San Francisco, CA, USA.

Gong, Z., Zhang, M., and Zhou, G. (2011). Cache-based Document-level Statistical Machine Translation. In *Proceedings of the 10th Conference on Empirical Methods in Natural Language Processing*, pages 909–919, Edinburgh, UK.

Gong, Z., Zhang, M., and Zhou, G. (2015). Document-Level Machine Translation Evaluation with Gist Consistency and Text Cohesion. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation*, pages 33–40, Lisbon, Portugal.

Gong, Z., Zhang, Y., and Zhou, G. (2010). Statistical Machine Translation Based on LDA. In *Proceedings of the 4th International Universal Communication Symposium*, pages 286–290, Beijing, China.

Gong, Z. and Zhou, G. (2015). Document-level Machine Translation Evaluation Metrics Enhanced with Simplified Lexical Chain. In *Natural Language Processing and Chinese Computing*, pages 396–403. Springer.

Guillou, L. (2013). Analysing Lexical Consistency in Translation. In *Proceedings of the 1st Workshop on Discourse in Machine Translation*, pages 10–18, Sofia, Bulgaria.

Guillou, L. and Hardmeier, C. (2016). PROTEST: A Test Suite for Evaluating Pronouns in Machine Translation. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B., and Popescu-Belis, A. (2016). Findings of the 2016 WMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the 1st Conference on Machine Translation*, pages 525–542, Berlin, Germany.

Guillou, L., Hardmeier, C., Smith, A., Tiedemann, J., and Webber, B. (2014). ParCor 1.0: A Parallel Pronoun-Coreference Corpus to Support Statistical MT. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, pages 3191–3198, Reykjavik, Iceland.

Halliday, M. A. K. and Hasan, R. (2014). *Cohesion in English*. Routledge.

Hamp, B. and Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. In *Proceedings of the Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, Madrid, Spain.

Hanks, P. (2000). Do word meanings exist? *Computers and the Humanities*, 34(1):205–215.

Hardmeier, C. (2014). *Discourse in Statistical Machine Translation*. PhD thesis.

Hardmeier, C. and Federico, M. (2010). Modelling Pronominal Anaphora in Statistical Machine Translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 283–289, Paris,France.

Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., and Cettolo, M. (2015). Pronoun-Focused MT and Cross-Lingual Pronoun Prediction: Findings of the 2015 DiscoMT Shared Task on Pronoun Translation. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation*, pages 1–16, Lisbon, Portugal.

Hardmeier, C., Nivre, J., and Tiedemann, J. (2012). Document-Wide Decoding for Phrase-Based Statistical Machine Translation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1179–1190, Jeju Island, Korea.

Hardmeier, C., Tiedemann, J., and Nivre, J. (2013). Latent Anaphora Resolution for Cross-Lingual Pronoun Prediction. In *Proceedings of the 12th Conference on Empirical Methods in Natural Language Processing*, pages 380–391, Seattle, WA, USA.

Hasler, E., Haddow, B., and Koehn, P. (2014). Dynamic Topic Adaptation for SMT using Distributional Profiles. In *Proceedings of the 9th Workshop on Statistical Machine Translation*, pages 445–456, Baltimore, MD, USA.

Heafield, K. (2011). KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland.

Hoek, J., Evers-Vermeul, J., and Sanders, T. J. (2015). The Role of Expectedness in the Implicitation and Explicitation of Discourse Relations. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation*, pages 41–46, Lisbon, Portugal.

Huang, S. F. (1995). Chinese as a Metonymic Language. In *In Honor of William Wang: Interdisciplinary studies on Language and Language Change*, pages 223–252, Taipei, Taiwan. Pyramid Press.

Iacobacci, I., Pilehvar, M. T., and Navigli, R. (2016). Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of the 54^{th} Annual Meeting of the Association for Computational Linguistics*, pages 897–907, Berlin, Germany.

Jean, S., Lauly, S., Firat, O., and Cho, K. (2017). Does Neural Machine Translation Benefit from Larger Context? *arXiv preprint arXiv:1704.05135*.

Jurgens, D. and Klapaftis, I. (2013). SemEval-2013 Task 13: Word Sense Induction for Graded and Non-Graded Senses. In *Proceedings of the 2^{nd} Joint Conference on Lexical and Computational Semantics and the 7^{th} International Workshop on Semantic Evaluation*, pages 290–299, Atlanta, GA, USA.

Kilgarriff, A. (1997). I Don't Believe in Word Senses. *Computers and the Humanities*, 31:91–113.

Kilgarriff, A. (2006). *Word Senses*. Springer Netherlands, Dordrecht.

Kingma, D. P. and Ba, J. (2015). Adam: a Method for Stochastic Optimization. In *Proceedings of the 3^{rd} International Conference on Learning Representations*, San Diego, CA, USA.

Knight, K. (1999). Decoding Complexity in Word-Replacement Translation Models. *Computational linguistics*, 25(4):607–615.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the 10^{th} Machine Translation Summit*, pages 79–86, Phuket, Thailand.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45^{th} annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180.

Koehn, P. and Knight, K. (2003). Empirical Methods for Compound Splitting. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics*, volume 1, pages 187–193, Budapest, Hungary.

Koehn, P. and Knowles, R. (2017). Six Challenges for Neural Machine Translation. *arXiv preprint arXiv:1706.03872*.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical Phrase-based Translation. In *Proceedings of the 4^{th} Conference of the North American Chapter of the Association for*

Computational Linguistics on Human Language Technology, pages 48–54, Edmonton, Canada.

Koskeniemmi, K. and Haapalainen, M. (1994). GERTWOL–Lingsoft Oy. *Linguistische Verifikation: Dokumentation zur Ersten Morpholympics*, pages 121–140.

Lapshinova-Koltunski, E. (2015). Exploration of Inter- and Intralingual Variation of Discourse Phenomena. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation*, pages 158–167, Lisbon, Portugal.

Le Nagard, R. and Koehn, P. (2010). Aiding Pronoun Translation with Co-Reference Resolution. In *Proceedings of the Joint 5th Workshop on Statistical Machine Translation and MetricsMATR*, pages 252–261, Uppsala, Sweden.

Lison, P. and Tiedemann, J. (2016). Opensubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, Portorož, Slovenia.

Loáiciga, S., Stymne, S., Nakov, P., Hardmeier, C., Tiedemann, J., Cettolo, M., and Versley, Y. (2017). Findings of the 2017 DiscoMT Shared Task on Cross-lingual Pronoun Prediction. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*, pages 1–16, Copenhagen, Denmark.

Luong, M.-T. and Manning, C. D. (2015). Stanford Neural Machine Translation Systems for Spoken Language Domains. In *Proceedings of the 13th International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Lyons, J. (1968). *Introduction to Theoretical Linguistics*. Cambridge University Press.

Mariño, J. B., Banchs, R. E., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A. R., and Costa-jussà, M. R. (2006). N-Gram-Based Machine Translation. *Computational Linguistics*, 32(4):527–549.

Martínez Garcia, E., Creus, C., España i Bonet, C., and Màrquez Villodre, L. (2017). Using Word Embeddings to Enforce Document-Level Lexical Consistency in Machine Translation. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*, pages 85–96, Prague, Czech Republic.

Martínez Garcia, E., España Bonet, C., and Màrquez Villodre, L. (2015). Document-Level Machine Translation with Word Vector Models. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation*, pages 59–66, Antalya, Turkey.

Martínez Garcia, E., España i Bonet, C., and Màrquez Villodre, L. (2014). Document-Level Machine Translation as a Re-translation Process. *Procesamiento del Lenguaje Natural*, 53:103–110.

Mascarell, L. (2017). Lexical Chains meet Word Embeddings in Document-level Statistical Machine Translation. In *Proceedings of the 3rd Workshop on Discourse in Machine Translation*, Copenhagen, Denmark.

Mascarell, L., Fishel, M., Korchagina, N., and Volk, M. (2014). Enforcing Consistent Translation of German Compound Coreferences. In *Proceedings of the 12th Konvens Conference*, pages 58–65, Hildesheim, Germany.

Mascarell, L., Rios, A., and Volk, M. (2016). Crossing Sentence Boundaries in Statistical Machine Translation. *MultiLingual*, pages 50–52.

Meyer, T. (2011). Disambiguating Temporal-Contrastive Discourse Connectives for Machine Translation. In *Proceedings of the 49th Annual Meeting on Association for Computational Linguistics Student Research Workshop*, pages 46–51, Portland, OR, USA.

Meyer, T. (2014). *Discourse-Level Features for Statistical Machine Translation*. PhD thesis.

Meyer, T., Hajlaoui, N., and Popescu-Belis, A. (2015). Disambiguating Discourse Connectives for Statistical Machine Translation. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 23(7):1184–1197.

Meyer, T. and Poláková, L. (2013). Machine Translation with Many Manually Labeled Discourse Connectives. In *Proceedings of the Workshop on Discourse in Machine Translation*, pages 43–50, Sofia, Bulgaria.

Meyer, T. and Popescu-Belis, A. (2012). Using Sense-Labeled Discourse Connectives for Statistical Machine Translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation and Hybrid Approaches to Machine Translation*, pages 129–138, Avignon, France.

Meyer, T., Popescu-Belis, A., Hajlaoui, N., and Gesmundo, A. (2012). Machine translation of labeled discourse connectives. In *Proceedings of the 10th Biennial Conference of the Association for Machine Translation in the Americas*, San Diego, CA, USA.

Meyer, T., Popescu-Belis, A., Zufferey, S., and Cartoni, B. (2011). Multilingual Annotation and Disambiguation of Discourse Connectives for Machine Translation. In *Proceedings of the 12th Annual SIGdial Meeting on Discourse and Dialogue*, pages 194–203, Portland, OR, USA.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the Workshop at the International Conference on Learning Representations*, Scottsdale, AZ, USA.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Morris, J. and Hirst, G. (1991). Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text. *Computational Linguistics*, 17(1):21–48.

Neubig, G. (2016). Lexicons and Minimum Risk Training for Neural Machine Translation. In *Proceedings of the 3rd Workshop on Asian Translation*, Osaka, Japan.

Novák, M. and Žabokrtský, Z. (2014). Cross-Lingual Coreference Resolution of Pronouns. In *Proceedings of the 25th International Conference on Computational Linguistics*, pages 14–24, Dublin, Ireland.

Och, F. J. (2003a). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, Sapporo, Japan.

Och, F. J. (2003b). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, volume 1, pages 160–167, Sapporo, Japan.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29(1):19–51.

Och, F. J., Ueffing, N., and Ney, H. (2001). An Efficient A* Search Algorithm for Statistical Machine Translation. In *Proceedings of the Workshop on Data-Driven Methods in Machine Translation*, pages 1–8, Toulouse, France.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, PA, USA.

Pelevina, M., Arefiev, N., Biemann, C., and Panchenko, A. (2016). Making Sense of Word Embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 174–183, Berlin, Germany.

Popescu-Belis, A., Meyer, T., Liyanapathirana, J., Cartoni, B., and Zufferey, S. (2012). Discourse-Level Annotation over Europarl for Machine Translation: Connectives and Pronouns. In *Proceedings of the 8th international conference on Language Resources and Evaluation*, Istanbul, Turkey.

Popović, M. (2017). Comparing Language Related Issues for NMT and PBMT between German and English. In *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*, pages 209–220, Prague, Czech Republic.

Pourvali, M. and Abadeh, M. S. (2012). Automated Text Summarization Base on Lexicales Chain and Graph using of WordNet and Wikipedia Knowledge Base (sic!). *Computing Research Repository.*

Pu, X., Mascarell, L., and Popescu-Belis, A. (2017). Consistent translation of repeated nouns using syntactic and semantic cues. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 948–957, Valencia, Spain.

Pu, X., Mascarell, L., Popescu-Belis, A., Fishel, M., Luong, N.-Q., and Volk, M. (2015). Leveraging Compounds to Improve Noun Phrase Translation from Chinese and German. In *Proceedings of the ACL-IJCNLP Student Research Workshop*, pages 8–15, Beijing, China.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann Publishers.

Rinaldi, A. M. (2009). An Ontology-Driven Approach for Semantic Information Retrieval on the Web. *ACM Transactions on Internet Technology*, 9(3):10.

Rios, A., Mascarell, L., and Sennrich, R. (2017). Improving Word Sense Disambiguation in Neural Machine Translation with Sense Embeddings. In *Proceedings of the 2nd Conference on Machine Translation*, Copenhagen, Denmark.

Rios Gonzales, A. and Göhring, A. (2013). Machine Learning Disambiguation of Quechua Verb Morphology. In *Proceedings of the 2nd Workshop on Hybrid Approaches to Translation*, pages 13–18, Sofia, Bulgaria.

Rios Gonzales, A. and Tuggener, D. (2017). Co-Reference Resolution of Elided Subjects and Possessive Pronouns in Spanish-English Statistical Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 657–662, Valencia, Spain.

Rothe, S. and Schütze, H. (2015). AutoExtend: Extending Word Embeddings to Embeddings for Synsets and Lexemes. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 1793–1803, Beijing, China.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation Methods for Unsupervised Word Embeddings. In *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal.

Sennrich, R., Firat, O., Cho, K., Birch, A., Haddow, B., Hitschler, J., Junczys-Dowmunt, M., Läubli, S., Miceli Barone, A. V., Mokry, J., and Nadejde, M. (2017). Nematus: a Toolkit for Neural Machine Translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain.

Sennrich, R. and Haddow, B. (2015). A Joint Dependency Model of Morphological and Syntactic Structure for Statistical Machine Translation. In *Proceedings of the 14th Conference on Empirical Methods in Natural Language Processing*, pages 2081–2087, Lisbon, Portugal.

Sennrich, R. and Haddow, B. (2016). Linguistic Input Features Improve Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation*, pages 83–91.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the 1st Conference on Machine Translation*, pages 368–373, Berlin, Germany.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725, Berlin, Germany.

Smith, K. S., Azizy, W., and Specia, L. (2016). The Trouble with Machine Translation Coherence. *Baltic Journal of Modern Computing*, 4(2):178.

Stairmand, M. (1996). *A Computational Analysis of Lexical Cohesion with Applications in Information Retrieval*. PhD thesis, University of Manchester.

Taghipour, K. and Ng, H. T. (2015). Semi-Supervised Word Sense Disambiguation Using Word Embeddings in General and Specific Domains. In *Proceedings of the 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 314–323, Denver, Colorado.

Tam, Y.-C., Lane, I., and Schultz, T. (2007). Bilingual LSA-based adaptation for statistical machine translation. *Machine Translation*, 21(4):187–207.

Thater, S., Fürstenau, H., and Pinkal, M. (2011). Word Meaning in Context: A Simple and Effective Vector Model. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 1134–1143, Chiang Mai, Thailand.

Tiedemann, J. (2010). Context Adaptation in Statistical Machine Translation Using Models with Exponentially Decaying Cache. In *Proceedings of the Workshop on Domain Adaptation for Natural Language Processing*, pages 8–15, Uppsala, Sweden.

Tiedemann, J. (2012a). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation*, pages 2214–2218, Istanbul, Turkey.

Tiedemann, J. (2012b). Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2214–2218, Istanbul, Turkey.

Tuggener, D. (2016). *Incremental Coreference Resolution for German.* PhD thesis, Universität Zürich.

Ture, F., Oard, D. W., and Resnik, P. (2012). Encouraging Consistent Translation Choices. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics. Human Language Technologies*, pages 417–426, Montréal, Canada.

Vilar, D., Xu, J., d'Haro, L. F., and Ney, H. (2006). Error Analysis of Statistical Machine Translation Output. In *Proceedings of the 5th Language Resources and Evaluation Conference*, pages 697–702.

Volk, M., Bubenhofer, N., Althaus, A., Bangerter, M., Furrer, L., and Ruef, B. (2010). Challenges in Building a Multilingual Alpine Heritage Corpus. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, Valletta, Malta.

Wang, L., Tu, Z., Way, A., and Liu, Q. (2017). Exploiting Cross-Sentence Context for Neural Machine Translation. *arXiv preprint arXiv:1704.04347.*

Williams, P., Sennrich, R., Nadejde, M., Huck, M., Haddow, B., and Bojar, O. (2016). Edinburgh's Statistical Machine Translation Systems for WMT16. In *Proceedings of the 1st Conference on Machine Translation*, pages 399–410, Berlin, Germany.

Wong, B. T. M. and Kit, C. (2012). Extending Machine Translation Evaluation Metrics with Lexical Cohesion to Document Level. In *Proceedings of the 2012 International Conference on Computational Linguistics on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1060–1068, Jeju Island, Korea.

Xiao, T., Zhu, J., Yao, S., and Zhang, H. (2011). Document-level Consistency Verification in Machine Translation. In *Proceedings of the 13th Machine Translation Summit*, pages 131–138, Xiamen, China.

Xiong, D., Ben, G., Zhang, M., Lv, Y., and Liu, Q. (2013a). Modeling Lexical Cohesion for Document-Level Machine Translation. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence*, pages 2183–2189, Beijing, China.

Xiong, D., Ding, Y., Zhang, M., and Tan, C. L. (2013b). Lexical Chain Based Cohesion Models for Document-Level Statistical Machine Translation. In *Proceedings of the 12th Conference on Empirical Methods in Natural Language Processing*, pages 1563–1573, Seattle, WA, USA.

Xiong, D. and Zhang, M. (2013). A Topic-based Coherence Model for Statistical Machine Translation. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, Bellevue, WA, USA.

Xiong, D. and Zhang, M. (2014). A Sense-Based Translation Model for Statistical Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1459–1469, Baltimore, MD, USA.

Zhang, R. and Ittycheriah, A. (2015). Novel Document Level Features for Statistical Machine Translation. In *Proceedings of the 2nd Workshop on Discourse in Machine Translation*, pages 153–157, Lisbon, Portugal.

Zhao, B. and Xing, E. P. (2006). BiTAM: Bilingual Topic Admixture Models for Word Alignment. In *Proceedings of the International Conference on Computational Linguistics and the Annual Meeting of the Association for Computational Linguistics*, pages 969–976, Sydney, Australia.

Zhong, Z. and Ng, H. T. (2010). It Makes Sense: A Wide-Coverage Word Sense Disambiguation System for Free Text. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. System Demonstrations*, pages 78–83, Uppsala, Sweden.

Ziemski, M., Junczys-Dowmunt, M., and Pouliquen, B. (2016). The United Nations Parallel Corpus v1.0. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 3530–3534, Portorož, Slovenia.